DOCUMENT RESUME

ED 167 843 CE 020 406

AUTHOR	Mullins, Cecil J., Ed.; Winn, William R., Ed.
TITLE	Criterion Development for Job Performance Evaluation: Proceedings from [a] Symposium (San Antonio, Texas, June 23-24, 1977).
INSTITUTION	Air Force Human Resources Lab., Brooks AFB, Tex. Personnel Research Div.
SPONS AGENCY	Air Force Human Resources Lab., Brooks AFB, Texas.
REPORT NO	AFHRL-TR-78-85
PUB DATE	Feb 79
NOTE	194p.; Some pages may not reproduce clearly due to light print
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	Aptitude Tests; Armed Forces; Conference Reports; Cost Effectiveness; *Criterion Referenced Tests; Data Processing; Employment Qualifications; *Job Skills; Military Personnel; Organizational Effectiveness; *Performance Criteria; Performance Specifications; Performance Tests; *Personnel Evaluation; Personnel Management; Problems; *Rating Scales; *Task Performance; Test Reliability; Test Validity; Work Sample Tests
IDENTIFIERS	United States

ABSTRACT

These conference proceedings contain statements from those concerned with various aspects of research on job performance measures. Following introductory remarks, the formal presentations of work and ideas include "Air Training Command Interest in the Criterion Problem," by Donald E. Meyer, "The Criterion Problem: A Personnel Management Perspective," by Wayne Sellman and Willibord T. Silva, "Army Research in the Criterion Area: A Change of Emphasis," by J. E. Uhlaner and others, "Navy Efforts in Criterion Development for Job Performance Evaluation," by Frederick A. Muckler, "The Criterion Problems: An Overview of Evaluation and Measurement Research in the AFHRL Technical Training Division," by Philip J. DeLeo and Brian K. Waters, "Overview of Advanced Systems Division Criterion Research (Maintenance)," by John P. Foley, Jr., "Criterion Problems," by Cecil J. Mullins and Forrest R. Ratliff, "Rater Accuracy," by Joseph L. Weeks and Cecil J. Mullins, rating research reports by Eric D. Curton, Cecil J. Mullins, and others, "Synthetic Criteria," by Cecil J. Mullins and others, and "What is the Value of Aptitude Tests?" by Raymond D. Christal. The report concludes with summary material provided by a panel of five research consultants from the civilian community who were invited to express their views on the work presented. (Author/CSS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

CE

ED167843

CE 020 406

AFHRL-TR-78-85

AIR FORCE HUMAN RESOURCES LABORATORY

CRITERION DEVELOPMENT FOR JOB PERFORMANCE EVALUATION:

PROCEEDINGS FROM SYMPOSIUM
23 AND 24 JUNE 1977

Edited by

Cecil J. Mullins
William R. Winn, SrA, USAF

PERSONNEL RESEARCH DIVISION
Brooks Air Force Base, Texas 78235

February 1979

Approved for public release; distribution unlimited.

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Personnel Research Division, under project 2313, with HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235. Dr. Cecil J. Mullins (PEP) was the Principal Investigator for the Laboratory.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved for publication.

LELAND D. BROKAW, Technical Director
Personnel Research Division

RONALD W. TERRY, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFHRL-TR-78-85	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (*and Subtitle*) CRITERION DEVELOPMENT FOR JOB PERFORMANCE EVALUATION: PROCEEDINGS FROM SYMPOSIUM 23 AND 24 JUNE 1977		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(*s*) Cecil J. Mullins William R. Winn		8. CONTRACT OR GRANT NUMBER(*s*)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Personnel Research Division Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2313T6
11. CONTROLLING OFFICE NAME AND ADDRESS HQ Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235		12. REPORT DATE February 1979 13. NUMBER OF PAGES 192
14. MONITORING AGENCY NAME & ADDRESS(*if different from Controlling Office*)		15. SECURITY CLASS. (*of this report*) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (*of this Report*)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (*of the abstract entered in Block 20, if different from Report*)

18. SUPPLEMENTARY NOTES

This report presents the proceedings of a symposium sponsored by the Air Force Office of Scientific Research (AFOSR) and hosted by the Air Force Human Resources Laboratory (AFSC).

19. KEY WORDS (*Continue on reverse side if necessary and identify by block number*)

aptitude tests
ipsative rating
job performance evaluation
performance measurement
personnel management
personnel rating criteria
personnel rating research

20. ABSTRACT (*Continue on reverse side if necessary and identify by block number*)

This report consists of the proceedings from a symposium conducted 23 and 24 June 1977 in San Antonio, Texas. The purpose was to bring together several of the researchers who have been recently concerned with various aspects of criterion research to exchange ideas over a 2-day period, and to provide discussion and critique of the directions our respective research efforts are taking. More formal presentations of work and ideas connected with criterion research by military scientists comprised the central part of the 2-day period. It was preceded by more informal material in the way of introductory remarks, and it was followed by summary material provided by a panel of five eminent researchers from the civilian community who were invited to serve as expert consultants and to give us their views concerning our work. The informal materials preceding and following the formal presentations were

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

Item 20 Continued:

taken directly from tape recordings of the proceedings, and, with minor editorial changes by the speakers (who were invited to review their remarks prior to publication) appear just as they were spoken.

SUMMARY

Air Force Human Resources Laboratory has recently launched an attack on the problems associated with producing a meaningful criterion measure of job performance. Changes in training technology are slowly destroying technical training performance as the criterion which historically has been used in the validation of selection and classification tests. This situation, of course, is decidedly inconvenient, but one healthy effect of it is that we are being forced to take a closer look at the possibility of developing a criterion more directly related to on-the-job performance, an effort which should continue across the years in any organization with a practical interest in predictor research.

We have high hopes, but few illusions. We know that the criterion problem has been perhaps the most intractable one in psychometrics since its inception. But we know also that, for some incomprehensible reason, few concerted and sustained efforts have been mounted on this most important research area. We do not expect to "solve" the criterion problem, but we hope we can make a few contributions, and we believe we can at least make some progress toward our modest goal--to develop a satisfactory substitute for technical school grades to use as a validation criterion for our predictor tests.

This symposium was sponsored by AFOSR, with the invaluable assistance of Captain Jack Thorpe. The purpose was to bring together several of the researchers who have been recently concerned with various aspects of criterion research to exchange ideas over a 2-day period, and to provide discussion and critique of the directions our respective research efforts are taking. More formal presentations of work and ideas connected with criterion research by military scientists comprised the central part of the 2-day period. It was preceded by more informal material in the way of introductory remarks, and it was followed by summary material provided by a panel of five eminent researchers from the civilian community who were invited to serve as expert consultants and to give us their views concerning our work. The informal materials preceding and following the formal presentations were taken directly from tape recordings of the proceedings, and, with minor editorial changes by the speakers (who were invited to review their remarks prior to publication) appear just as they were spoken.

We sincerely hope that the publication of these proceedings will be representative of the most advanced thinking currently available on criterion research. We confidently believe that this publication contains thinking which will be helpful to anyone directly concerned with this challenging and fascinating area.

6

PREFACE

We are pleased to express our appreciation to all the participants in the symposium who worked so hard on the papers presented here, and we offer our special thanks to the five invited members of a panel requested to offer criticism and guidance to the rest of us. They were, in alphabetical order:

Dr. John P. Campbell
Psychology Department
University of Minnesota
Minneapolis MN 55455

Dr. Richard J. Campbell
AT&T
Basking Ridge NJ 07920

Dr. Robert M. Guion
Psychology Department
Bowling Green State University
Bowling Green OH 43403

Dr. John S. Helmick
Educational Testing Service
Princeton NJ 08540

Dr. Ernest J. McCormick
Department of Psychological Sciences
Purdue University
W. Lafayette IN 47907

PROGRAM

SYMPOSIUM ON CRITERION DEVELOPMENT FOR JOB PERFORMANCE EVALUATION

23 through 24 June 1978

Thursday, June 23

Hidalgo Room	Morning Session Chairperson Col T. H. Newton Air Force Human Resources Laboratory
0815 - 0820	Opening Statement Dr. Charles E. Hutchinson Air Force Office of Scientific Research
0820 - 0830	Air Force Human Resources Laboratory Welcome Col Dan D. Fulgham Air Force Human Resources Laboratory
0830 - 0835	Administrative Announcements
0835 - 0850	Keynote Address Dr. Leland D. Brokaw Air Force Human Resources Laboratory
0850 - 0920	Air Training Command Interest in the Criterion Problem Dr. Donald E. Meyer
0920 - 0950	The Criterion Problem: A Personnel Management Perspective Maj Wayne Sellman Air Force Military Personnel Center
0950 - 1010	Coffee Break
1010 - 1145	Army Research in the Criterion Area: A Change of Emphasis Dr. J. E. Uhlaner U.S. Army Research Institute for the Behavioral and Social Sciences
1145 - 1155	Administrative Announcements
1155 - 1330	Lunch

Hildago Room	Afternoon Session Chairperson Dr. Nancy Guinn Air Force Human Resources Laboratory
1330 - 1415	Navy Efforts in Criterion Development for Job Performance Evaluation Dr. Fred Muckler Navy Personnel Research and Development Center
1415 - 1500	The Criterion Problem: An Overview of Evaluation and Measurement Research in the AFHRL Technical Training Division Dr. Philip J. DeLeo Air Force Human Resources Laboratory
1500 - 1510	Break
1510 - 1555	Overview of Advanced Systems Division Criterion Research Dr. John P. Foley, Jr. Air Force Human Resources Laboratory
1555 - 1640	Criterion Problems Dr. Cecil J. Mullins Air Force Human Resources Laboratory
1640 - 1700	Discussion
1700 - 1705	Administrative Announcements

END THURSDAY SESSION

Friday, June 24

Hildago Room	Morning Session Chairperson Dr. Lonnie D. Valentine, Jr. Air Force Human Resources Laboratory
0815 - 0845	Rater Accuracy Joseph L. Weeks Air Force Human Resources Laboratory
0845 - 0945	Rating Research 1. Content Analyses of Rating Criteria Capt Eric D. Curton 2. The Normative Use of Ipsative Ratings Dr. Cecil J. Mullins
0945 - 1005	Coffee Break

1005 - 1050	Synthetic Criteria Lt Col Forrest R. Ratliff Air Force Human Resources Laboratory
1050 - 1150	What is the Value of Aptitude Tests? Dr. Raymond Christal Air Force Human Resources Laboratory
1150 - 1200	Administrative Announcements
1200 - 1330	Lunch
Hildago Room	Afternoon Session Chairperson Dr. Leland D. Brokaw Air Force Human Resources Laboratory
1330 - 1500	Comments by Consultants, Roundtable Discussions by Consultants and R&D Representatives
1500 - 1540	Closing Comments and Adjournment

END OF SESSIONS

TABLE OF CONTENTS

Page

I. Opening Statement . . . . Dr. Charles E. Hutchinson 1

II. Welcoming Remarks . . . . .Col Dan D. Fulgham 3

III. Introduction to Keynote Speaker . . . . . .Col T. H. Newton 5

IV. Keynote Address . . . . .Dr. Leland D. Brokaw 6

V. Air Training Command Interest in the Criterion Problem . . . . Donald E. Meyer 9

VI. The Criterion Problem: A Personnel Management Perspective . . . . Wayne Sellman 13
Willibord T. Silva

VII. Army Research in the Criterion Area: A Change of Emphasis . . . . J. E. Uhlaner 19
A. J. Drucker
W. B. Camm

VIII. Navy Efforts in Criterion Development for Job Performance Evaluation . . . . Frederick A. Muckler 39

IX. The Criterion Problem: An Overview of Evaluation and Measurement Research in the AFHRL Technical Training Division . . . . .Philip J. DeLeo 57
Brian K. Waters

X. Overview of Advanced Systems Division Criterion Research (Maintenance) . . . . John P. Foley, Jr. 68

XI. Criterion Problems. . . . . .Cecil J. Mullins 98
Forrest R. Ratliff

XII. Rater Accuracy. . . . . Joseph L. Weeks 110
Cecil J. Mullins

XIII. Rating Research
1. Content Analyses of Rating Criteria . . . . .Eric D. Curton 116
Forrest R. Ratliff
Cecil J. Mullins
2. The Normative Use of Ipsative Ratings. . . . . Cecil J. Mullins 123
Joseph L. Weeks

Page

XIV. Synthetic Criteria. . . . . Cecil J. Mullins 128
Forrest R. Ratliff
James A. Earles

XV. What is the Value of Aptitude Tests?. . . . . .Raymond D. Christal 131

Summary and Conclusions 146

Consultant Comments 147

Summary Statements. 160

Impressions 173

Comments on Symposium on Criterion Development for Job Performance Evaluation . . . . John S. Helmick 174

Comments from the Sidelines . . . . Ernest J. McCormick 177

12

I

OPENING STATEMENT

Dr. Charles E. Hutchinson
Air Force Office of Scientific Research

I have a memory for all of the wrong things. I can remember one time spending 10 weeks in San Antonio, and the reason for being here was to deactivate the Air Force Personnel and Training Research Center. Some of you may have memories that long. My role was to cull through the productive efforts of a lot of people both in-house and by contractual support in the area of social psychology and social sciences, which was supposedly my field, and recommend which should go to the archives, which should go to the burn basket, and which to try to salvage.

And I can bet you that this is a much happier time to be in San Antonio to not bury Caesar but to praise him, and it's been one of the delights of my short career in OSR--I've only been there since 1956, the same year that I deactivated AFPTRC--and I got hooked by OSR and it became an addiction.

But the reason for OSR being involved is that OSR is a research arm of the Air Force which reaches out to the research community in universities. For your information, I think in the year to come, 1978, and the years following on, there will be an enhanced Air Force research program in universities, and AFOSR will be the key instrument for the Air Force in reaching the universities with this program. I simply tell you that to alert you. Many of you are in service, some of you may by that time be out, but don't forget OSR. It's a place that will be available. The new research program is being sponsored by the Department of Defense. I can tell you what the planning was when I was a part of the system, and it was that the first year would be 33 million dollars, 11 million in each of the services for expanded university defense research, the second year would be 50 million with whatever proportion would go equally to the services, and the third year a 75 million dollar program, 25 million in each of the services.

Now if this program comes to OSR (and they're still talking about it--Dr. Allen and Dr. Gomoda are still in place), we're going to need some help in encouraging people to do meaningful research that has justification for the Air Force--not for the National Science Foundation, not for the National Institute of Health--and it's OSR's role to manage a program of this kind which includes university research and other

13

research organizations working for the Air Force to assure that this is coupled with the needs both current and future of Air Force laboratories. The prime laboratory that I have been concerned with and for which I'm most grateful because they have made it easy to do my coupling job is the Human Resources Laboratory through its divisions. It is another evidence of that coupling that I'm here today and that OSR can have a small part in fostering a program that invented the concept of having a meeting. The work was done here in the Personnel Division, and I'm able to take all this credit simply because there was a concept in OSR to expend some resources in trying to improve the coupling, and OSR's been at that point.

I'd like to make one introduction. I'm here talking for OSR as if I belonged. It's correct that I am a retired person and not a program manager anymore; I'm almost a free citizen. I've got under two weeks, I think, to finish this year's quota that they've allotted me. But Capt Jack Thorpe is the official and substantial representative of AFOSR--you may have known him as a substantial member of the Flying Training Division program--but he will be with us and he is the program manager in the area in which this meeting operates. So if you have ideas and you want to sell somebody, don't tell me, tell him. Jack will be fomenting this program to the best of his abilities, and we are convinced in OSR that they're substantial. I really, as I said, have nothing to say other than welcome and get with it.

II

WELCOMING REMARKS

Colonel Dan D. Fulgham
Commander
Air Force Human Resources Laboratory

It's a great pleasure for our laboratory to host this meeting. I came down here with some intention of making a few opening remarks and remind you of the importance of this kind of work, but seeing the people in the audience--I think I probably know 90% of you--and since this isn't Sunday, there's no sense in me preaching to the choir today. I would like to welcome you and tell you I believe that, as psychologists, you're in very good hands. Ty Newton's a physiologist; Dr. McCormick will tell you that I'm more physiologist than psychologist, so we think we can probably do you a good turn. But we are very pleased to have you here.

Charley made some remarks in connection with the demise of personnel research except for the small unit that we had left at Lackland. When I came into the organization back in 1971, I started asking questions about why should the work that apparently was so important to the Air Force have fallen into enough disfavor of support that we actually wound up losing a considerable organizational capability. I think Charley, if I'm correct, you went from about twelve hundred people down to 800 and finally wound up with about 250 left at Lackland when they disestablished the organization. And I think that probably one of the major reasons that led to the lack of support at the higher management levels of the organization was that the research efforts got too far from the user requirements. It seemed that when it was time for the user to stand up and be counted and support the laboratory, he couldn't find enough usable research that was being directly applied to some of his problems. I think that probably one of the things that we have to guard against in this business more than anything else is the production of useful but not used research.

Now we've taken a new tack in this laboratory in that we try to ensure that when we start working on a user problem, he is convinced it's a problem, that we share that conviction, and we try and draw him into our research with us. And I think that that has paid off enormously for us in that we're getting a better pickup on our product than ever before. Now, since I'm principally experienced in the flying end of the business, we, of course, have been very, very much interested in research, over time on the performance of the pilots and aircrews. I was reminded by a colleague from the University of Michigan recently that we've been working on objective performance measurement for 30 years in flight regimes and we're no closer to having a viable system than we were when we started. So, something that I think you'll be hearing about today--hopefully you'll mention it--is the pilot skills maintenance program that we're trying to generate. We're trying to draw a lot of this human performance under an umbrella program that we're going to call Skills Maintenance and Reacquisition Training. Now a key element of this--step number 2 after the identification of the skills in which we're principally interested--is the measurement of performance in those skills. And hopefully, for the first time (and we have some indication we may be successful this time), we're going to convince the Air Force to let us scientifically or technically manipulate these skills and their performance and measure the effects. From this, hopefully, will come the data base that we need. Then we need to determine what kinds of training programs, what combinations of media, and what kind of a training system we need in the aircrew area. I think there'll be a great deal of spin-off from this into the other areas of performance measurement as well.

III

INTRODUCTION TO KEYNOTE SPEAKER

Colonel Tyree H. Newton
Chief, Personnel Research Division
Air Force Human Resources Laboratory

I mentioned earlier that in order to get something like this off the ground it takes a lot of people doing a lot of things. The prime mover for this symposium was Dr. Leland Brokaw. It was his idea. He discussed it over a year ago and it kind of faded for awhile, and then he brought it up again, and he kept with it. He's the one who made the contact with Dr. Hutchinson, he provided the theme and the format for this symposium, and it's through his persistence that we're here today. Dr. Brokaw has been with this organization, or the precursor of this organization, since 1946 as a civilian. Prior to that time he was with it for 3 years in the military, so he knows the business. He's held virtually every type of job in personnel research and he's presently the Technical Director for the Personnel Research Division. It's with pleasure that I introduce to you Dr. Leland Brokaw, who will give the keynote remarks for this symposium.

IV

KEYNOTE ADDRESS

Dr. Leland D. Brokaw
Technical Director
Personnel Research Division
Air Force Human Resources Laboratory

Col Fulgham warned about preaching to the choir and I find myself in that somewhat unenviable position, but it seemed to me that a few comments to perhaps set the tone for this meeting would be in order. I realize a keynote speech is supposed to arouse your passions and your enthusiasms, and we all go forward to defeat the foe and all those good things, so this really isn't a keynote; this perhaps is more of a footnote. In passing, I'd like to point out that numbers of us have heard an announcement proffered by my friend, Fred Muckler, who is back there in the bleachers someplace. The Navy is having a similar kind of meeting focused on their problems in performance measurement, October 12 through October 14, in San Diego, and I look forward to being there. It is our hope that some of the things that are perhaps conceived here will be born there.

We are met to discuss a basic problem in personnel management. We are met to discuss an intractable difficulty in personnel research. We are met to discuss an area in which there has been scientific frustration and lack of confidence for many, many years. Yet in a pragmatic world of work we see businesses, industries, and military services going about their missions in productive ways with apparent happiness on the part of the people who work for them. So why then are we making such a big deal of developing ways of objectively measuring performance on a job? Is it because we lack the ingenuity, is it because we do not perceive the true complexity of work environments, or is it because we are making the job too complicated for ourselves? Col Fulgham supported us in October of 1976 when we launched a program in criterion development. He knows that we know that the probability of our finding a glorious solution is relatively small. He knows, as we know, that if we do find such a solution, it will be to the considerable benefit of most industries, most industrial psychologists, most organizations.

Our goal is to develop a methodology for the collection of job performance data for use in the validation of Air Force selection and classification devices. It's parochial, it's narrow, and it's our problem; it's the one we want to talk about here today.

There are three reasons we want to do this: First, changes in training technology are slowly destroying technical training performance as our criterion to be used in the validation of selection and classification tests. If we look at pass/fail we find that the PQ splits are 90 to 10 or worse. Air Training Command has recognized our problem. They are continuing to develop a continuous numeric score for many of the courses at some cost to themselves.

Secondly, we have recognized ever since I started this business, longer ago than most of you have been here, that the technical training grade as a device for the validation of a selection instrument is an interim kind of criterion. The objective of selection, like the objective of training, is to put a competent worker in a job. While it is true the completion of training is a hurdle that you must get by to get to the job, there is as yet very little demonstration of relevance of the selection or the training for the job. We must generate a system that will permit the judgment of such relevance.

The third reason was forecast in my opening comments. A research problem exists here, ad hoc developments for the purpose appear in the literature by the thousands, but there does not appear to be a continuity, a flow, which establishes systems which can be applied objectively by comparatively untrained people which will generate useful data for our purposes. Assessment centers for the identification of managers or the pinpointing of places where managers need training are very popular these days. We thought about assessment centers for perhaps 45 seconds and concluded that the ponderous nature of the time that they take and the amount of money that they cost renders them undesirable as useful measures for the validation of enlisted selection measures in the Air Force. An eminent psychologist, whose name I can't remember, has contemplated this problem and he has said, "It's going to cost you a lot of money to collect performance data to use for a criterion. But be that as it may, if that's what it costs, go ahead and spend it." Well, these are nice, brave words for a guy who doesn't have my budget.

In our own program, our approach has been classical. I'm afraid we've shown very little ingenuity. We're starting from all the well known places. But it is our intent by doing this to tie together the shreds we find in the literature and to build a basis for further progress. We've always got an eye on the checkbook. It is our intent to balance costs to get results. If we are completely successful, we'll have a straightforward, inexpensive, objective way of collecting the kind of data that we need.

Now you all know that there are performance measuring systems operational in every organization for every kind of people in these organizations. But there are differences between those kinds of data and the kind that we need for the validation of classification devices. We need devices that are sensitive to individual differences in job specific skills. If it's possible, we need to measure those skills in a way that is uncontaminated by the personality and the motivations of the incumbent. At the same time we need also to measure that motivation, the drive, the initiative, so that we can moderate, if you will, the aptitude data that we collect. The performance evaluations used in operational programs tend to be more generalized; they tend to be overall measures of productivity or performance. They tend to be focused on promotability rather than on the things which make the current job really well done or not well done. And, we have another problem. Insofar as a supervisor cannot hire or fire or promote unilaterally, insofar as a supervisor is not culpable for high ratings, insofar as a supervisor depends upon his people for his own production, there will be a tendency for him to rate high. When ratings get high they lose their variance, and when they lose their variance they lose their predictive efficiency. We find this in most military performance programs.

This conference has three major objectives. First, to share our areas of concern and difficulty, that we may jointly explore for economic solutions. Secondly, to review ongoing efforts in the Personnel Research Division for the elicitation of constructive criticism. Thirdly, to foster common attacks upon our common problems, the best approach to this business. With the experience and the expertise provided in this group, we'll have a better chance than we've ever had before to really begin to cope with some of the basic issues of this matter. Let us move into the presentations of this symposium with an awareness of the difficulties of the area, with confidence that there are ways to solve them. Let us be critical in our search for effective techniques, and let us be alert for the positive things in every presentation that we'll hear.

20

V

AIR TRAINING COMMAND INTEREST IN THE CRITERION PROBLEM

Donald E. Meyer
Air Training Command
Randolph Air Force Base, Texas

The main theme of this symposium has to do with performance criteria as they apply to personnel selection and classification, and you may be assured that the Air Training Command has vital and continuing interests in these areas. But after the selection and classification process is completed, the Air Training Command is faced with providing the most effective and economical training possible. Consequently, in recognition of our extended interests, Dr. Brokaw gave me permission to change the thrust of my presentation to the need for performance criteria for training purposes.

As many of you know, the Air Force has been committed to the use of instructional system development (ISD) since about 1970, first by policy statements from the Air Force Chief of Staff, and more recently by Air Force regulation. Additionally, conceptual guidance is given in Air Force Manual 50-2, and "How To" information for application of ISD to course development is provided by Air Force Pamphlet 50-58. An ISD'ed course is based on the exact requirements of the specialty for which the training is provided. It is a key to the avoidance of unnecessary and therefore wasteful training. Avoidance of waste has always been important to skillful and conscientious course developers, but now becomes a necessity due to budgetary restraints.

Although the Air Training Command led the Air Force in the use of ISD in course development, we are still beset with many problems. Better training for ISD practitioners is a continuing need. Additionally, ISD training for management personnel needs to be further emphasized to make management more aware of the time, effort, and resources that must be invested in a really first-class ISD treatment; and, of course a realization of the efficiencies that result, i.e., teaching precisely what is needed for the job. These are real problems, but solutions come readily to mind and there is hope that if not by edict, perhaps through osmosis they will be solved over time.

The biggest problem and the one for which I can see no near term solution lies in the early phases of applying the ISD process, the task analysis. In addition to being the first step in the ISD process, it is also the most crucial, for without the proper data base, expressed in usable detail, the effort rests on a bad foundation. The result, though perfectly executed, will likely fall short of providing the most cost-effective training possible, i.e., it may teach either more or less than the skills required on the job. The likelihood is that the course will contain more than required, and that is wasteful. Non-ISD believers scoff at this idea by asserting that no one can ever know too much. I agree with them in principle, but the notion assumes that having once been exposed to a skill or subject matter in a school situation, it is retained for application at some later time. This premise seldom holds true. Again, what is needed is an accurate and reliable means to identify the performance requirements of the job. In theory we know how to do this, but in practice some elements are missing. We do not have access to task analyses for most of the skills we train. And with an obligation to conduct some 3,000 different courses, of which about one-third are revised each year, it is doubtful that we will ever have conventional task analyses for this purpose. Our budget simply won't accommodate this expense. Let me explain how we presently do business, what the constraints are, and what needs to be improved.

One of the prime documents used in course development is the specialty training standard (STS). This is an Air Force publication used to standardize and control the subject matter content and level of training perceived as needed to achieve the skills and knowledge required for an Air Force specialty. It is prepared by the particular ATC school responsible for the training and then circulated through the major Air Force commands for review and coordination, after which it is published to become a quasi-contract between ATC as the producer and the MAJCOMs who receive our graduates.

The STS is a widely used document. It has been around for about 25 years or so and has wide acceptance in the Air Force. It provides a listing of the knowledges and skills that should be possessed for an Air Force specialty and, as such, it provides a start point in the development cycle. The STS is used as a basis for resident course development, OJT, follow-on career development courses, and other functions such as development of the specialty knowledge tests which are used for promotion considerations. It is a useful document, but it does have several limitations that should be given a great deal of attention.

The first and most obvious is the fact that the STS is developed by subject matter specialists who rely on their own backgrounds and experience to determine what it should contain. I can't knock experience--it's a valuable asset--but frequently people with similar experience backgrounds have entirely different views on the same topic. Also, even though the people who develop the STSs bear the same AFSC, some of them have had different experiences during their careers and this also leads to disagreements. Who is right? The outcome is usually arbitrary, but predictably represents the views of the highest ranking, most articulate, or vociferous member of the team developing the STS. Errors made are generally on the conservative side and that's why the MAJCOMs don't take issue with an STS during coordination. The training is seen as adequate even though it might be of wider scope and depth than would actually be required. We have had a lot of help on this particular problem, based upon AFHRL research in improving the efficiency of our occupational survey techniques. I'd like to briefly summarize some things that are happening that are encouraging to the belief that the STS can be made more objective than it now is. Periodically, the Occupational Measurement Center, an ATC organization, conducts occupational surveys. All of the enlisted AFSCs in the Air Force with authorizations of over 100 personnel in an occupational specialty are surveyed. This occurs at about 3- to 4-year intervals. An exhaustive listing of duties and tasks for a particular specialty is developed by a group of senior and knowledgeable personnel in each specialty gathered from MAJCOMs Air Force-wide. The listing is then put into a survey format and sent to the field where performance data are gathered. Prior to the AFHRL research in this area, occupational survey reports resulted in voluminous machine printouts and addressed only the number of airmen performing the tasks and the percent of time they spent on them. Though they provided reliable data, these printouts proved tedious to analyze and incomplete for use in curriculum development. Course designers still had to base their decisions on many undefined subjective factors such as "task criticality," "task importance," etc.

The recently developed product of HRL research promises to virtually automate the decision making process. The research has identified and quantified the major factors of the previously subjective judgments. These new factors, task delay tolerance, consequences of inadequate performance, and task difficulty can be statistically combined with the old factors to yield a training priority index. This index ranks each task in a specialty in the order of its priority for training. From these data, a fairly objective picture of what people in the field are actually doing and the implications for training can be obtained. The Command has recently developed a procedure that uses the occupational survey data to construct specialty training standards. At present, the procedure is being service tested at several of our technical training centers. If the present service test proves the technique successful, a big obstacle, that is, the subjectivity of the STS will have been overcome. This will give us a certain amount of assurance that the STS is based upon actual field requirements rather than what someone thinks those requirements are.

Even with this improvement, however, the STS task items are too broadly stated to be used in the development of behavioral objectives for efficient training. For example, in one of the electronics career field STSs, a task statement says "Align the system." This is an important maintenance function and it is simple and understandably stated. Upon a closer look, however, we find that there are some 50

23

alignments that can be made on a given piece of equipment. You can readily see the dilemma faced in trying to apply ISD with that kind of imprecise data base. The STS task segments are just not specific enough. The course developer is forced to exercise subjective judgments that can be very wasteful in terms of over-training or dangerous in terms of under-training.

What we need is a method that will translate the task statements of the STS into task analysis-type detail usable for course development. The process must be reliable, fast, and economical. I have seen a classification of nine different approaches to task analysis. This classification ranges all the way from on-site observation to a single subject matter expert making a detailed break-out of task data. Each of these approaches has its advantages and disadvantages. The most reliable approach, i.e., on-site observation by a skilled analyst, is prohibitively expensive; the least expensive approach, the subject matter expert, is too prone to personal bias to be creditable. The solution we seek must exist someplace between these extremes at a point where we could sacrifice an acceptable percent of reliability for a great enough reduction in cost to make the process affordable.

We need the help of the research community in the development and validation of a technique or techniques to solve this problem. The training establishments of the services would be the most immediate beneficiary, but there are other applications as well: the production of job performance aids, the production of maintenance instructions for technical orders and perhaps, since the task analysis data we need for training is closely related to the performance data needed for the development of improved selection assignment techniques, it might be possible for a contribution in this area. I would urge that you keep this is mind as you shape your research programs. The refinement of present task analysis techniques or a breakthrough in finding a new approach that would result in economical and reliable task data in sufficient detail to be used in course development is sorely needed and will require at least as great a research effort as was expended in the improvement of the STS.

VI

THE CRITERION PROBLEM: A PERSONNEL MANAGEMENT PERSPECTIVE

Major Wayne S. Sellman and Lt Col Willibord T. Silva
Air Force Military Personnel Center
Randolph Air Force Base, Texas

Within the Air Force, we are confronted with the same personnel problems as any other organization, whether large or small, public or private--that of shaping and adapting available human resources into useful and effective manpower. In that regard, the very multiplicity of skills required by the Air Force poses problems in personnel planning, training, and manpower utilization which are all but unprecedented. Personnel requirements change rapidly and on a large scale, and are dependent to a large extent upon technological advances and the international political situation.

Obviously, Air Force personnel management is a highly complex affair. As you know, to cope with these complexities requires creative and innovative personnel research--research which addresses all aspects of the personnel life-cycle: selection, classification, training, performance appraisal, promotion, and organizational development. Such topics are of great interest to us--an interest engendered from two basic sources. First, we are users of your product. Our effectiveness as personnel managers hinges on the successful application of techniques and procedures developed from past personnel research.

Second, we are sponsors of your research. In that role, we serve as the liaison agency between you and the rest of the Air Force encouraging, explaining, and extolling the virtues of research and its applications.

Thus, we have a very symbiotic relationship with personnel research scientists. We depend on you for timely and efficient solutions to management problems as well as for input into the formulation of personnel policy. You, in turn, depend on us as sort of public relations experts who ensure your various efforts are understood and appreciated not only across the Air Force rank and file but at the highest echelons of Air Force management as well. So, we were especially pleased to accept the invitation to speak at this symposium and share some of our ideas and perceptions with you.

Now, to the subject at hand. We were asked to comment on the Air Staff interest in the criterion problem. That interest can be expressed in one word--considerable; in fact, to overstate its importance to personnel management would be literally impossible. How we do business in personnel is to a large extent determined by the criteria used in personnel research. Without adequate criteria, personnel functions derived from and dependent upon that research would be less effective and efficient. In other words, the magnitude of the contribution of personnel research to Air Force personnel management is determined, for the most part, by the adequacy of the criterion measures evolved.

Having now established our interest in the criterion problem, perhaps it would be appropriate for us to identify just what we mean by a criterion. Blum and Naylor (1968) define criterion as a "measure of the goodness of a worker." Don't we wish this were so in the Air Force? In industrial personnel research, the criterion that is usually used concerns the degree to which a worker can be considered successful on the job. For example, the criterion might be sales figures, numbers of acceptable units produced, or any other measurement of work accomplishment, or lack thereof. Unfortunately, in the Air Force we have no overall measure of job success or productivity although one has been sought for the last 35 years.

Other definitions of the criterion may also be found in the literature. Guion (1965) defines it simply as "that which is to be predicted," while McCormick and Tiffin (1974) have described it in terms of "a dependent variable." It would seem that the Air Force rather pragmatically subscribes to these latter two definitions. In practice, our primary criterion is success in training; its rationale is that if a person is adequately trained, he will have sufficient knowledge to be able to successfully perform his job.

Although much work on the criterion problem has been accomplished, especially in measuring success in training, perhaps the time has come to shift emphasis and explore other types of criteria--criteria such as attitudes, motivation, satisfaction, leadership, accidents, absenteeism, and rates of promotion. Take the latter two, for example. All other things being equal (and they almost never are) the employee who attends work regularly is more valuable to the organization than the one who frequently misses work. If patterns of absence could be reliably measured, they might serve to open a new dimension in military selection research.

Moreover, even though the Air Force uses a weighted factor promotion system for enlisted personnel, length of time before promotion occurs, or number of times considered before promotion selection might be measures of promotability that could be used. Admittedly, because of constraints unique to the Air Force, such criteria may not be as easily measured and possibly not as directly relevant as if they were industrial criteria. Nevertheless, perhaps more attention should be directed toward their possible use. And, of course, there is still our old friend, job productivity. Even though past efforts haven't exactly yielded a breakthrough, pursuits in this direction must be continued.

Recently, selection research in the military services has been criticized by the Defense Science Board as well as other committees and working groups chartered by the Office of the Director, Defense Research and Engineering, for apparent lack of progress. These groups point out that validities are no higher today, on the average, than they were a decade ago. It is commonly accepted, although not necessarily by testing researchers, that the reason for this situation lies in the types of tests that are used as predictors (i.e., we have reached the state-of-the-art). However, another equally likely explanation may be in the way in which the criterion problem has been handled. Psychologists have traditionally sought "the criterion." To do that we have attempted to combine several subcriteria into one overall measure of job performance. But, as we have become more sophisticated, we have moved toward a position that job success is multidimensional in nature. If this is so, then it would follow logically that criteria must also be multidimensional. Could it be that one way to enhance our selection and classification strategies would be through the use of multiple criteria? Too often, we do not use all the job information available in the selection of criteria. True, time and cost considerations come into play, but more effort should be expended in selecting criteria appropriate for each individual military occupation, not just using success in training as the catchall criterion for all of them.

In this regard, we believe that one of the best statements of this point was made by Wallace and Weitz in the 1955 Annual Review of Psychology: "The criterion problem continues to lead all others in lip service and to trail most in terms of work reported. It seems probable that almost all investigators now recognize the importance of developing acceptable criteria and submitting them to the greatest scrutiny and correction. Unfortunately, a reviewer must also conclude that the pressure of getting things done is still wooing many into the convenient device of accepting the criteria at hand and hoping it will turn out all right." Unfortunately, this situation is even today, some 20 years later, still the rule rather than the exception.

Now one final word about the selection of criteria. Brogden and Taylor (1950) have identified ten major criterion problems encountered by personnel researchers. One of these is sponsor acceptability--the selection of a criterion that is meaningful and fully acceptable to management. We would suggest that today's researchers, particularly those in the military environment, are not as sensitive to this consideration as they could and should be. For example, in planning studies, how often do scientists interact with research users in the selection of criteria. Probably not very often. A more common occurrence might be the scientist selecting the criteria and then informing the user--if even that much coordination goes on in the research planning stages. Clearly, here is an area where research can be made more user

27

oriented--the user must be involved in the selection of "acceptable and relevant" criteria.

The issue of relevance introduces an area of criterion technology alluded to earlier, i.e., operational/mission effectiveness. Using the best criteria available, we have selected, classified, and trained a highly capable personnel force and sent them to the field with assurances to commanders that these people can do the job. What now? How does the commander know that the job is being done, or, even more importantly, that the mission will be accomplished when or if the horn blows? Every commander is seeking that evasive assessment of organizational effectiveness which represents the operationalization of the skills and capabilities of his personnel.

Typically, we in the military have assessed overall mission effectiveness in terms of the four factors shown in Figure 1. For the combat unit all of these are relevant; for support units different combinations of the four factors are more appropriate. However, regardless of the unit's mission or function one factor remains constant--personnel.

We make our evaluations of the non-personnel factor in fairly quantitative terms using computer modeling, engineering tests, combat experience, and on-site inspections. Our assessment of the human factor is much less sophisticated. War games or exercises and operational inspections are our typical tools, but these are subjective at best as well as time constrained. When we consider that in a year's time 20% of a unit's personnel may have changed, the effectiveness rating received 12 months earlier takes on an entirely different perspective. Thus, the requirement for quantifiable, integrated, time-sensitive criteria for organizational effectiveness remains a technology need.

The literature on organizational effectiveness criteria is extensive and, because of its ubiquitousness, has made application difficult and somewhat limited. While organizational criteria have been described in terms of system input/output/process variables, identification of potential standards alone is not enough. Such identification must be followed with the development and validation of reliable and relevant criteria of organizational effectiveness. Bowser, in an August, 1976, review concerning criteria of operational unit effectiveness, summarizes the requirement quite succinctly: "The basic problem of defining organizational effectiveness within the U.S. Navy (all Services) requires considerable research. The framework established for evaluation of criteria is general enough to fit most organizational criteria. However, because it is so general, it may not provide sufficient structure for evaluation. The state-of-the-art concerned with evaluating organizational effectiveness is primitive enough to require development of criteria in order to support organizational research."

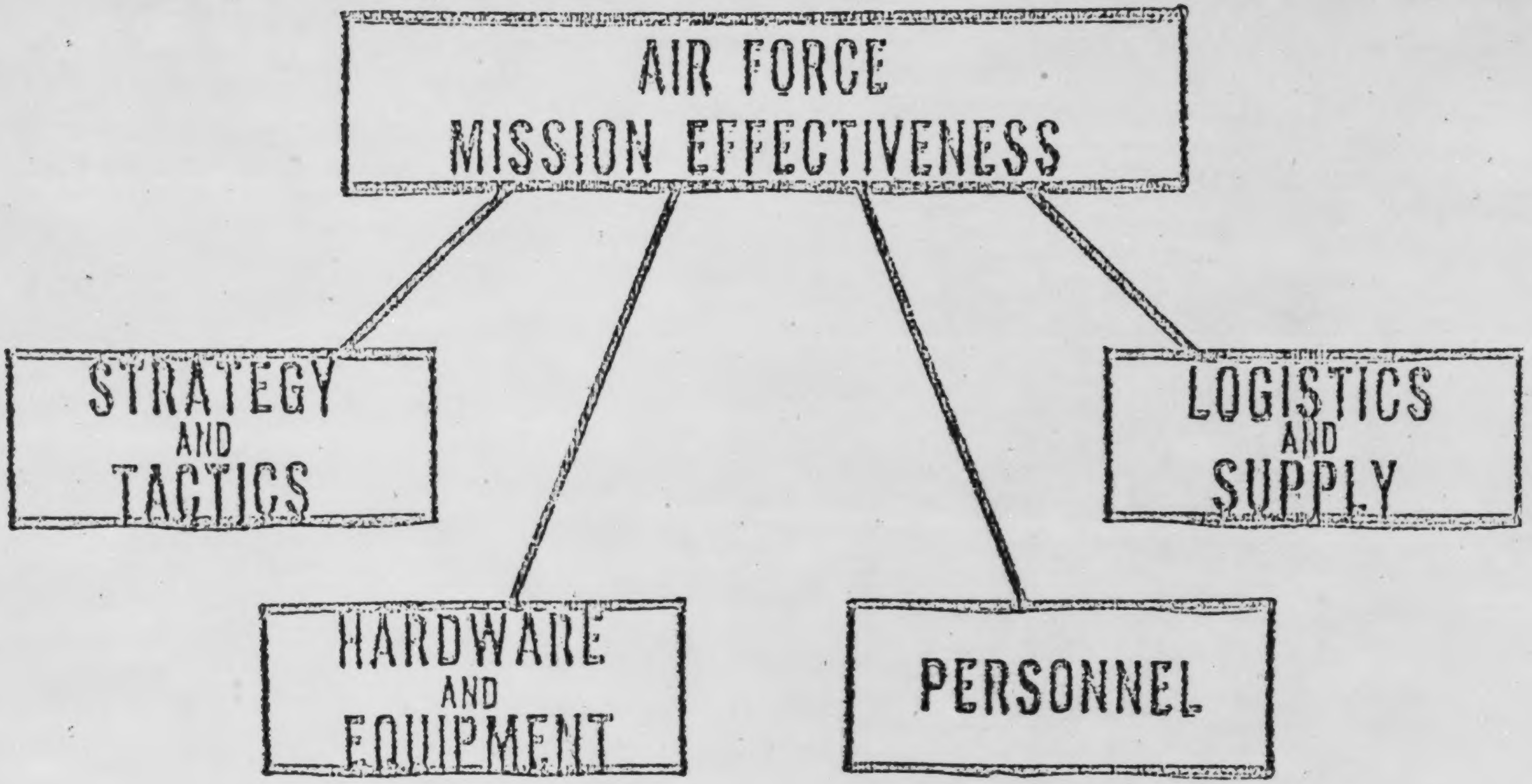

AIR FORCE MISSION EFFECTIVENESS
STRATEGY AND TACTICS
HARDWARE AND EQUIPMENT
PERSONNEL
LOGISTICS AND SUPPLY

Figure 1. Factors Related to Mission Effectiveness

Our latter excursion into organizational effectiveness was obviously not intended to provide a learned treatise on operational criteria technology. It was, rather, designed to sensitize you to a legitimate user need. We must not forget that the personnel pipeline extends far beyond its input junction. Indeed, perhaps its reach beyond that point represents the most challenging albeit most rewarding advancement of criterion technology.

In summary, Air Staff interest in criteria is to find the best one(s), combine them in the most appropriate and imaginative way, and accordingly streamline to the maximum extent the way we do business in "hiring, placing, progressing, and evaluating" our people. However, as Blum and Naylor (1968) have pointed out, "For years, psychologists have labored under the notion that the objective is to find 'the criterion' in the same way that the knights of King Arthur's Round Table were charged with finding the Holy Grail. Both have had about equal and limited success." We trust that in the ensuing 9 years this situation has somewhat improved. Certainly Patricia Cain Smith (1976) in her chapter on criteria in the Handbook on Industrial and Organizational Psychology sounds a note of optimism. In any event, development of reliable, relevant, and valid criteria for use in Air Force personnel research (and management) remains a task of paramount importance. It's nice to be present at this symposium and to know there are the kinds of people represented here who are capable of addressing this difficult problem.

REFERENCES

Blum, M.L., & Naylor, J.C. *Industrial psychology: Its theoretical and social foundations*. New York: Harper and Row, 1968, Chapter 7.

Bowser, S.E. *Determination of criteria of operational unit effectiveness in the U.S. Navy* (NPRDC-TR-76TQ-41). San Diego CA: Navy Personnel Research and Development Center, August 1976.

Brogden, H.E., & Taylor, E.K. The dollar criterion: Applying the cost accounting concept to criterion construction. *Personnel Psychology*. 1950, 3, 133-154.

Guion, R.M. *Personnel testing*. New York: McGraw-Hill, 1965, Chapter 4.

McCormick, E. J., & Tiffin, J. *Industrial Psychology* (6th ed.). Englewood Cliffs NJ: Prentice-Hall, Inc., 1974, Chapter 2.

Smith, P.C. Behaviors, results, and organizational effectiveness: The problem of criteria. *Handbook of Industrial and Organizational Psychology*. (M.D. Dunnette, Ed). Chicago: Rand McNally College Publishing Company, 1976, Chapter 17.

Wallace, S. R., & Weitz, J. Industrial Psychology. *Annual Review of Psychology*, 1955, *6* 217-250.

VII

ARMY RESEARCH IN THE CRITERION AREA:
A CHANGE OF EMPHASIS

H.E. Uhlaner, A.J. Drucker, and W.B. Camm
U.S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

During the past decade, Army research to develop and measure criteria for human performance has moved to achieve greater relevance to job tasks, including the noncognitive aspects of these tasks and more efficient implementation of performance measures related to Army problems. That is, criteria are expected not only to be psychometrically predictable but to show reasonably logical, relevant relationships to the job. There is wide recognition that few job performances are unidimensional, also an awareness that it is neither possible nor feasible to test completely all the component tasks and subtasks of many jobs or work situations. Hence critically selective sampling plans have been developed. Information concerning how well an individual can perform the tasks necessary to do the job is often gathered by means of a "criterion reference test"--a test made up of items directly related to the job of interest (Boycan & Rose, 1977). Adequate and relevant statistical measurement of job performance is either not practical or rigorous; often influenced by noncognitive considerations, e.g., degree of risk taking. New assessment indicators had to be developed and used along with more conventional methods. Analytic experience has convinced the performance test community that there is no easy way to overcome chronic criterion validity problems. Only meticulous, knowledgeable development of accurate descriptions of the relationships between psychological variables and precise identification of these variables can reduce criterion validity problems. The minimal passing criterion, the way this criterion was derived from the job objectives, the nature of the test items, and the length of the test together make up the assessment system, within which a variety of quantitative models are used (Macready, Steinheiser, Epstein, & Mirabella, in press).

The Test Bed Model

For a better understanding of job performance criteria it has become very clear that a better theoretical base is necessary. The senior author has presented a concept of the interaction of selection, training, and job design for effective work performance. His major hypothesis is that aptitudes, job demands, and surrounding conditions coalesce to yield varying levels of performance. The conceptual background for his hypothesis includes a job taxonomy containing cognitive variance and noncognitive variance, the ad hoc nature of values and goals, and the great variety of styles of behavior by which individuals and organizations seek and achieve goals (Uhlaner, 1970).

It is proposed that for many applied purposes, including systems development, the criterion should be a given one, rather than the yield of preceding predictors, and should be explicitly specified with respect to both cognitive and noncognitive variance.

Figure 1 presents a test bed model which can be developed at the user's location. The user can indicate specifications of the results he desires. He is provided with a number of negotiable options leading to the same result, each reflecting a different trade-off possibility. The user makes the final decision as to the option selected (Uhlaner, 1970).

The test bed model method emphasizes the outcomes of decisions and their consequences for individuals and institutions, whereas traditional assessments have emphasized only measurement and prediction. The validity coefficient tells us about the degree of association between the predicted and obtained criterion scores. But often, from a practical standpoint, the number of correct personnel decisions resulting from the use of a given cutoff score is more important than knowledge of the validity coefficient (Cronbach & Gleser, 1965).

Achievement Criteria

Army Research Institute for the Behavioral and Social Sciences' (ARI) research results over the decades show that, in general, three types of criteria are used to measure achievement: school grades, ratings, and situational or performance measures. The trend, to no one's surprise, has been away from grades and more subjective ratings toward multi-criteria performance-oriented measurement. Table 1[1] shows the relative frequency with which these criteria occur in reports of ARI research over a 20-year period.

Table 1. Type and Frequency of Criteria Used
(N = 209 Publications, 1956 - 1977)

Type of Criteria		f
I.	Grades	79 (27%)
II.	Ratings	81 (27%)
III.	Performance	93 (31%)
-	Multi-Criterion	43 (15%)
		296

33

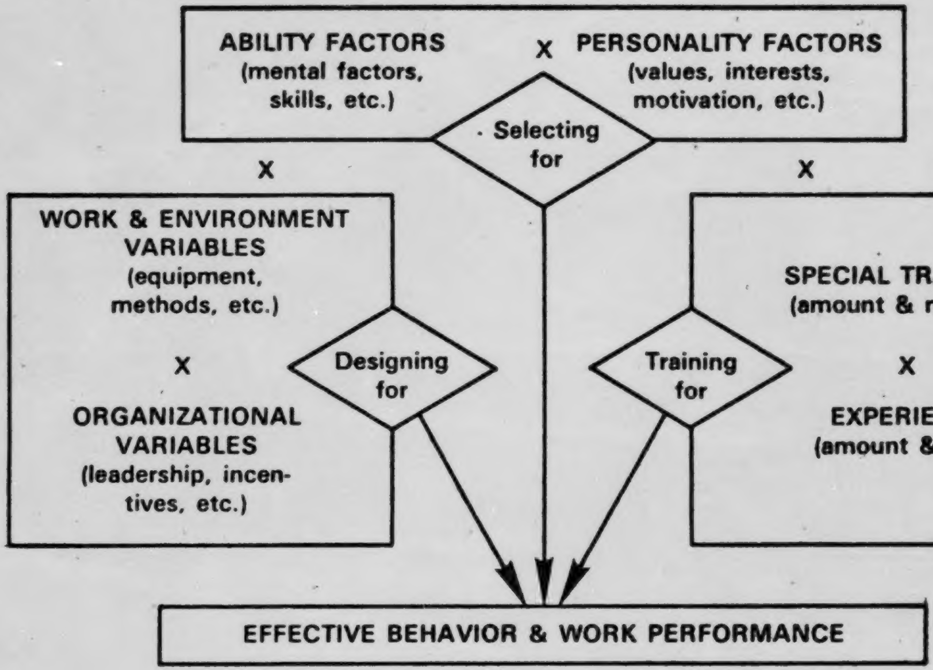

ABILITY FACTORS
(mental factors, skills, etc.)
X
PERSONALITY FACTORS
(values, interests, motivation, etc.)
Selecting for
X
X
WORK & ENVIRONMENT VARIABLES
(equipment, methods, etc.)
X
ORGANIZATIONAL VARIABLES
(leadership, incentives, etc.)
Designing for
Training for
SPECIAL TRAINING
(amount & method)
X
EXPERIENCE
(amount & type)
EFFECTIVE BEHAVIOR & WORK PERFORMANCE

Figure 1. Conceptualization of interactions of human factor system variables as related to performance effectiveness.

Grades are used primarily as criteria for cognitive predictors. Cognitive factors are those that involve acceptable right and wrong answers or job elements (Uhlaner, 1970). Grades are used as criteria for selection and classification tests, much the same as in the past (Haggerty, 1953; Maier, 1972; Zeidner, Harper, & Karcher, 1956). The recently implemented Skill Qualification Testing System (Maier, Young, & Hirshfeld, 1976) will gradually replace the paper-and-pencil Military Occupational Specialty (MOS) tests in the Army, however, thus reducing even further the need for grades. Ratings have been used to evaluate on-the-job performance of officers and enlisted men, especially where interaction with other people is involved. Selected performance tests have been used primarily to measure a more complicated mix of cognitive and noncognitive job demands.

The three groupings of criteria are not mutually exclusive and are intended only to provide some indication of the framework of their use--particularly within the ARI. Note that grades and ratings account for a little over half (54%) of the criteria used. This is due, in large part, to the larger proportion of studies involving school criteria. Also, current trends, as mentioned before, show that training and other performance criteria are increasingly obtained from simulated or situational performance-oriented indices.

Grades

By far the most frequently used criterion in the period just following World War II was the academic grade or the pass-fail training criterion. The relationship between grades and on-the-job performance has consistently never been very high. Yet where school training is a prerequisite for job assignment, the trainee must pass the course, and therefore the applied research scientist must pay some attention to grades or pass/fail measures in training. School grades appear to predict best when training is for jobs with high cognitive demands that involve clear-cut "right" and "wrong" job elements. Validity coefficients tend to be moderate to high between such jobs and school grades. In sum, grades are most useful in reflecting ability in academic or cognitive aspects of the job.

Grades in school do not seem to take into account noncognitive factors that relate to style of behavior and performance reflecting specified or implied values and attitudes. Experience on the job seems to be most crucial for specific noncognitive performance--experience coupled with the person's use of his/her individual talents and values to achieve goals.

Ratings

The rating is one measure of effectiveness that seems widely accepted. The essence of a rating is a judgment by one person or a group of persons of the performance of another individual. The rating is simple and familiar, but it is also the source of many fallacious beliefs among management and supervisors. ARI research for many years has attempted to establish methods for obtaining reliable and valid ratings; it has had its impact on many research tasks. However, many of the fallacies prevailing in the 50's are still with us. Here are some examples together with research-based information bearing upon the problem:

Fallacy 1. We can always meaningfully rate a person's performance on 30 to 40 separate scales. Research results have shown that a large general factor dominates the rating even when deliberate attempts are made to measure different aspects of job performance by using a number of specific rating scales. Raters typically seem to perceive only a single measure of success, whether it is an actual single measure, a formally weighted composite, or an implicit weighted composite. However, recent efforts to develop performance criteria have the practical advantage of combining related fractional criteria into a composite, tending to avoid the ambiguity of combining unrelated variables. This procedure defines related performance measures that are more clearly understood by the evaluators (Duffy, 1976; Root, Epstein, Steinheiser, Hayes, Wood, Sulzen, Burgess, Mirabella, Erwin, & Johnson, 1976). Criterion measures that assess individual job performance in terms of concrete job functions seem to yield a reasonably accurate measure of performance, whether or not the measures are subsequently combined into a composite rating. Also, multiple evaluators are likely to increase the validity of performance ratings.

Fallacy 2. Hard raters render more valid ratings than easy raters. In research addressing this subject, there is very little difference in validity of hard and easy ratings, although hard raters tend to bunch their ratings somewhat lower on the scale (Browning, Campbell, Birnbaum, Campbell, Fold, & Haggerty, 1952a, 1952b).

Fallacy 3. Bright raters render more valid ratings than the not-so-bright, or a rater has to be exceptionally bright to rate well. The research evidence is that raters of average intelligence have rendered ratings as valid as any rating by others. There is some evidence that when persons in the lower 16% of the distribution of mental abilities rate others, the ratings are not quite so valid (Chesler, Brogden, Brown, & Katz, 1952). However, nearly all raters tend to evaluate good performance more effectively than poor performance.

Fallacy 4. A better rating can be obtained by giving the rater a more definite frame of reference. An example of this would be "How would you like the ratee to serve under you?" rather than "How competent is the ratee?" The earlier research answer was that if any improvement results, it was negligible (Karcher, Campbell, Falk, & Haggerty, 1952). However, when measures are behavioral in content and actually relate to the expected behavior and the criterion dimensions underlying such measures are clearly identified, then reliable construct measurement techniques are effective.[2] The work in this area is still under way, and problems with the many theoretical aspects of current concepts of content and construct validity are moot. In any case, raters seem to rate more reliably and validly when they are aware of the criterion to be evaluated.

On investigation, thus, these four commonly held concepts have not proved to be entirely correct. However, several questions are often asked about rating practices and procedures that affect the research usefulness of the rating. Typical questions and answers in connection with the Officer Efficiency Rating are: Should every military officer be required to show his rating to the rated officer? It should make very little difference whether the ratings are shown or made by identified or anonymous raters, provided all ratings are done the same way (Chesler, Brogden, Brown, & Katz, 1952; Karcher, Winer, Falk, & Haggerty, 1952; Seeley & King, 1956). Are ratings by identified raters any different from ratings by anonymous raters? The consensus is that although there may be an inflation of ratings when the ratings are shown, differences in validity are negligible. Do raters agree more on their evaluations of job success if they have had more opportunity to observe the individual performing on the job? The answer is yes, generally, as implied in Table 2 (Medland & Olans, 1964).

Table 2 also shows superior validity of peer ratings, which have proven to be generally reliable and valid over cadre ratings (Mohr, 1975). One can reason that fellow trainees or fellow workers on the job are usually in a good position to observe performance, and that frequent association in a training situation, even for a period of 8 weeks, is sufficient to enable the rater to make the judgments required.

Table 3 shows some of the research evidence for the claim that the peer rating is one of the best predictors of subsequent Army performance (Downey, 1976; Drucker, 1957; Parrish & Drucker, 1957; Willemin, Rosenberg, & White, 1957).[3]

Table 3. Peer Rating Comparisons

Combat	r. = .60
Leadership	r. = .49
Special Forces	r. = .43
West Point	r. = .50
Ranger	r. = .52

Another important finding in most rating situations is that a rating based on the judgment of more than one rater is better than a single rating (Karcher et al., 1952). The use of multiple raters is quite likely to increase the validity of the performance rating.

TABLE 2

RELATIVE PREDICTIVE EFFICIENCY OF 4th AND 8th WEEK PEER AND CADRE RATING AND ADVANCED INFANTRY TRAINING (AIT) EXPERIMENTAL AND CONTROL SAMPLE

	AIT RATINGS			
	AIT EXPERIMENTAL PERSONNEL		AIT CONTROL PERSONNEL	
	PEER (N = 61)	CADRE (N = 49)	PEER (N = 69)	CADRE (N = 15)
PEER RATING 4th WEEK	.63	.27	.67	.85
PEER RATING 8th WEEK	.68	.42	.65	.75
CADRE RATING 4th WEEK	.34	.23	.11	.10
CADRE RATING 8th WEEK	.31	.23	.09	.14

Note. From Medland and Olans, 1964.

However, evaluations with different organizational perspectives are likely to yield different validity measures of the individual ratee's performance. More information is obtained, resulting in an even more accurate and possibly more useful assessment of performance (Duffy, 1976). It is the authors' conviction that ratings should be used most frequently when the assessment of noncognitive factors is involved, as in the performance of potential leaders or the performance of fighting personnel.

In sum, ratings are seen as simple to understand and easy to use. But ratings permit only relative measurements between person A and person B. For go/no go measurement, we must consider the third type of criterion--performance measurements.

Performance Measures

This third measure of effectiveness is one of the oldest and also, as one of the newest, has become increasingly acceptable.[4] In principle a performance test is a job sample test--similar in form to the trade test of the early years in industrial psychology. The test of performance in an actual situation has been applied with growing frequency where the need for more objective measures is perceived as crucial.

The advantages of the situational performance measure make it a much more effective criterion measure than the grade or rating, even though the development of such measures presents challenging problems. With performance tests, we can approach success/failure limits--a goal not reachable with traditional ratings. For example, how many hand grenades can the soldier throw on target in one minute? Or, how long does it take a squad to capture a specified hill? With such precise information, a commander can better assess the performance of individuals or groups; with ratings such comparison is less feasible because the needed reference point is lacking.

REALTRAIN. A most effective use of performance testing is exemplified in REALTRAIN, one of the Army's new and extremely successful tactical training systems (Root et al., 1976). The measurement objectives of REALTRAIN include a specific set of operations for observing and evaluating agreed-upon relevant kinds of behavior. The recorded data indicate whether or not a clearly operationally-defined job or task has been performed. The soldier's performance is measured directly--no inference is necessary. Simulated battlefield realism is an important consideration, so the performance objectives for combat effectiveness require that:

(1) Leaders and soldiers take timely and appropriate response to enemy action in a dynamic combat situation.

(2) Units achieve effective and efficient intra- and inter-unit coordination.

(3) Units maximize the effects of available weapons on the enemy.

(4) Units minimize the effects of enemy weapons on themselves.

The REALTRAIN method provides realism for two-sided, free-play exercises, with a credible means of assessing casualties. Infantry REALTRAIN exercises are centered around the M16 rifle. Each soldier's weapon is equipped with a 6X telescope (Fig 2), and all participants wear 3½" black two-digit numbers on their helmets. Opponents try to read each other's numbers using the telescope. When a man on one side identifies a number, he fires a blank round and reports the number to a controller; the controller then radios the number to a controller with the opposing force, and the man whose number was identified is assessed as a casualty[5] (Shriver, Griffin, Jones, Word, Root, & Hayes, 1975). Procedures have been developed to determine casualties objectively for the M-60 machine gun, hand grenade, M18A1 Claymore mine, LAW, tank main gun, TOW, DRAGON, and M16A1 antipersonnel and M-21 antitank mines.[6] A critical element of the tactical engagement simulation occurs during the after-action review, when events surrounding each day's action are discussed and feedback is provided each individual involved in the exercise.

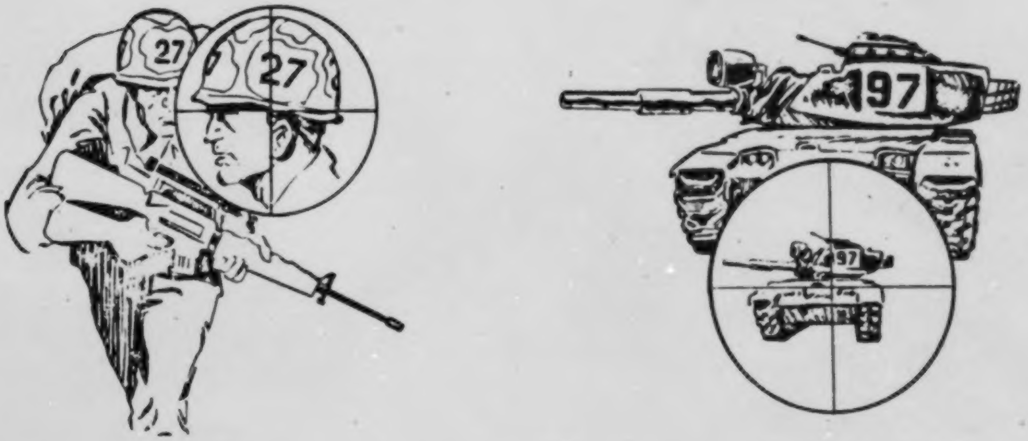


Figure 2. REALTRAIN simulation identification.

REALTRAIN is based on two conceptual frameworks. The first, as outlined by Uhlaner (1970), specifies human performance in systems terms; the second is based on the premise of the performance situation, in this case "success in battle." The initial validation of REALTRAIN (Root et al., 1976) with Army combat units in Europe and validation research at Fort Ord, California (Banks, Hardy, Scott, Kress, & Word, 1977), have indicated that training effectiveness results are impressively and consistently positive.

An obvious disadvantage of such performance measures or situational tests, however, is that they are difficult and expensive to construct. Despite efforts to facilitate the administration of standardized job elements, the observer's task remains a demanding one. Whenever possible, ARI relies on automatic recording of responses. One example, related to REALTRAIN, is the Multiple Integrated Laser Engagement Simulation Systems (MILES) (Fig 3), a family of low power, eye-safe lasers which will simulate the direct fire characteristics of the M16A1 rifle, the M60, M2, and M5 machine guns, the VIPER, DRAGON, TOW, and Shillelagh missile systems plus the 105mm tank main guns. A hierarchy of weapons effects is established in the detector logic--for example, a tank main gun can destroy an armored personnel carrier, but an M16 rifle cannot. This equipment provides immediate and accurate casualty assessment in two-sided, free-play tactical exercises.[7] The laser "firings" are keyed by the discharges of a blank round. Despite the sophisticated apparatus, a knowledgeable official is still needed to ensure that proper procedures are followed.[8] Thus, a need still exists to train observers thoroughly and rehearse them repeatedly in what they are to do.

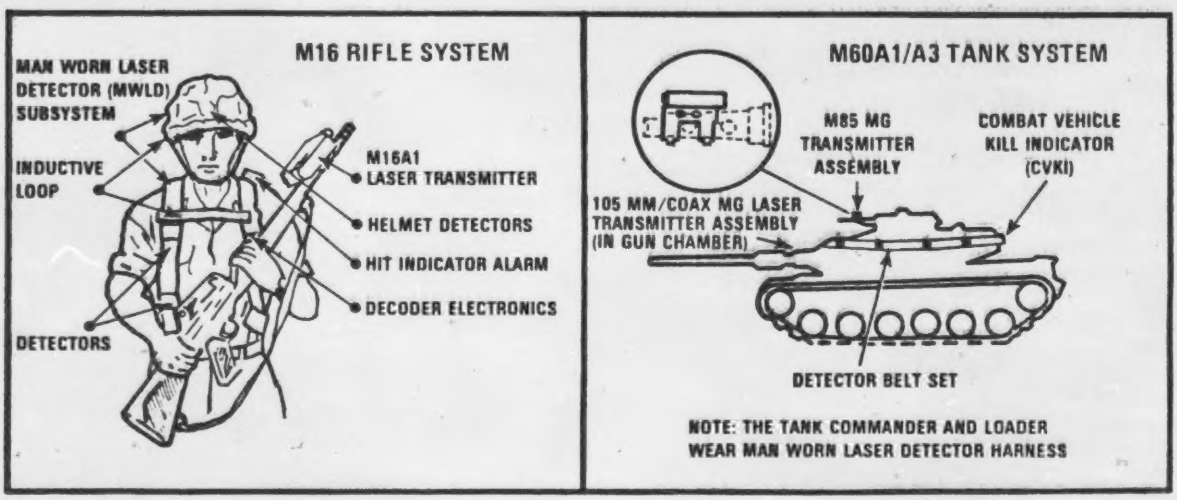

M16 RIFLE SYSTEM
MAN WORN LASER DETECTOR (MWLD) SUBSYSTEM
INDUCTIVE LOOP
DETECTORS
M16A1 LASER TRANSMITTER
HELMET DETECTORS
HIT INDICATOR ALARM
DECODER ELECTRONICS
M60A1/A3 TANK SYSTEM
M85 MG TRANSMITTER ASSEMBLY
COMBAT VEHICLE KILL INDICATOR (CVKI)
105 MM/COAX MG LASER TRANSMITTER ASSEMBLY (IN GUN CHAMBER)
DETECTOR BELT SET
NOTE: THE TANK COMMANDER AND LOADER WEAR MAN WORN LASER DETECTOR HARNESS

Figure 3. Multiple Integrated Laser Engagement Simulation (MILES).

Organizational effectiveness. A somewhat different area of measurement deals with the diagnosis and evaluation of Organizational Effectiveness (OE), often requiring situational performance measures of a largely non-cognitive nature--especially measures of attitudes and values.[9] The Work Environment Questionnaire (WEQ), used on OE research, provides attitude measures of the supervisors and the work group, gives situational factors that are related to job performance, and relates their importance to the job as perceived by the soldier and his leaders. The WEQ has been validated against objective standards of job activity and self-perceptions of work, all of which were in turn validated against actual on-the-job performance (Turney & Cohen, 1976).

The objective of the OE program is to identify and to optimize those organizational factors in the Army work environment related to soldier job satisfaction, motivation, and performance. The objective is met through a five-phase research program, progressively identifying and developing:

(1) Criteria of organizational effectiveness.

(2) Organizational functioning: structures, processes, and problems.

(3) Parameters of the OE process.

(4) Diagnostic methods.

(5) Intervention strategy.

The WEQ study was a follow-up of extensive longitudinal research encountered over a 3-year period to develop the diagnostic instruments. Pretests in 1973 provided initial data, validation of the instruments was conducted in 1974 and 1975, and in May-June 1975, an original diagnostic survey was conducted in one Army agency in the Army Air Defense Command. The survey focused primarily on Morse operations in a field station. Experimental considerations were:

(1) The work was performed by 16-man teams, each consisting of a senior NCO supervisor in charge of 14 operators and one analyst.

(2) Both individual and team performance criteria could be collected for validation purposes while the team did its job.[10]

(3) A large number of teams performing identical functions allowed experimental control.

The Morse operations are important to the mission requirements of the organization and the representation of the complex semicomputerized systems being implemented Army-wide (Cohen & Turney, 1976).

The findings, in general, revealed seven major organizational problem areas: peer group norms which fail to encourage good performance, insufficient performance feedback, need for training in supervisory technique, role ambiguity and conflict, inadequate intergroup communication patterns, lack of clear performance-reward relationship, and ambiguous performance evaluation standards. OE intervention was able to alleviate most of these.

Duty modules. An example of the development of performance criteria is the duty module concept which has the practical advantage of a composite criterion combining related variables that operationally define performance measures to the evaluators. The duty module is a cluster of tasks that are meaningfully related though not necessarily contained in one job. In fact, an ARI research project found that eight job dimensions could be incorporated into a single Job Proficiency Appraisal instrument designed to assess 30 entry-level specialty fields of the Officer Personnel Management System. These job dimensions describe specific duties in the areas of Administrative Details, Correspondence, Counseling, Maintaining Standards, Training, Supply Management, Technical Knowledge, and Control/Coordination (Duffy, 1976).

NOE. Situational performance tests demand both subject matter expertise and psychological knowledge. Imagination and ingenuity are required to bring out the desired performance in a highly concentrated test behavior simulation, contrived and presented for the examinee within limited geographical bounds. A host of practical problems must be solved. One example of a field problem is that used by Army helicopter performance evaluators.

The helicopter pilot's task is to navigate or fly a UH-1 helicopter over a prescribed route at Nap-of-Earth, or tree top height, at variable air speeds, using natural features for concealment. The performance is conducted in the field, and three measures are used.

(1) Total mission flights - a distance/track deviation measure which tells the percent of track followed and to what degree the pilot has been off course.[11]

(2) Individual tasks - tasks abstracted from total performance, such as mission planning (Farrell, 1973).

(3) Special individual behaviors - a high degree of abstraction is often involved here and, for that reason, the measurement of such behaviors is most readily accomplished in the laboratory. For example, levels of ambient illumination can be varied in order to determine effects upon terrain recognition ability.[12]

Besides the practical complications in measuring performance in the complex and multidimensional task of pilots, there is the problem of weight in the value of an error (e.g., the operational significance of a course deviation error of 300 meters, versus a deviation of 50 meters). This is a typical problem presented by performance measures that are tied to operation missions.

Despite these practical difficulties, a strong belief exists among performance research scientists in the human factors area that further progress in more sophisticated differential validation of certain kinds of human factors performance, particularly the kinds to which future officers of the Army may be exposed, can best be tapped by this sort of field/laboratory measurement. Earlier we implied that ratings hit only a common core of ability. We believe that situational performance measures will permit a sharper delineation of differential ability, as already evidenced by the Fort McClellan research project on officer performance.

Peculiar to the military and to the Army, whatever criteria are used, is the fact that jobs must be performed under both peacetime garrison and combat conditions. One of the biggest challenges has been how to secure effective measurement of performance in the combat situation. Combat situations are relatively rare, of course, and, when we find them, it may be extremely inconvenient to secure complete evaluations. Recognizing the importance for military psychologists of obtaining measures against such elusive combat criteria, research scientists have developed an approach called criterion equivalence (Wherry, Ross, & Wolins, 1954). The fundamental procedure in criterion equivalence approaches is based on a mathematical truism, that when two measures are equal to a third, they are equal to each other. Criterion equivalence studies have led to the conclusion that the same measures are predictive of performance in both combat and in garrison situations. The specific techniques of accomplishing criterion equivalence are elaborated in reports by Gaylord (1953) and Johnson (1956).

Systems Criteria

Underlying the discussion thus far have been the concepts of comparing one person with another, or one person against a specific set of job standards. As our laboratories have become concerned with systems and system research, we have become more aware of the fact that the systems the Army will be required to manage have very complex internal structures, and that if we are to learn how to act so as to produce the results intended, we will need new ways of thinking about complex systems (Uhlaner, 1960, 1964, 1975).

Development of the systems output criterion has proved to be somewhat more difficult. The generalized concepts that the military manager or system developer intuitively intended are very difficult to translate into operational terms. Systems evaluations are primarily a matter of

44

judgment by experts; and the larger the system, the more complex and difficult the translation from concept to operation becomes. Because of side effects and contingencies, many of the tasks do not have the outcomes intended. One of the greatest challenges for systems psychologists is to develop meaningful tasks that carry out system objectives.

From a situation where man has been the focal point, he has now become a linkage in a system. These systems are also becoming more and more expensive not only in dollars but in time lag. For any particular military function--for example, Command and Control--a number of competitive man-machine systems are being developed on a concurrent basis, and they have to be evaluated before they become operational. The evaluation of these competitive systems must be sound enough to enable military managers, together with the scientists, to make correct decisions as to the appropriate system or subsystem to be carried to completion or made operative.

The research psychologist has been asked to assist in establishing the appropriate subsets of functions to be performed--the jobs of the men within the chosen system. He is asked to indicate the kind of people needed, not only in terms of talents and aptitudes, but also, where appropriate, even in terms of personality characteristics. The researcher is asked to establish interrelationships and hierarchies within the system, to look at equipment and help engineers to design it, in order to make functions and jobs easier and more manageable by the average person. Concurrently, he is asked to develop training programs and devise aids which will, in the time allotted, train each individual to perform these functions. He is asked to look at the activities performed by the individuals after their training to see whether he can improve work methods. In the meantime, in theory, the machines will have been frozen in their design. In practice, all the processes of development are recycled many times. It is the last contingency that makes human factors problems more fluid, more complicated, more of a challenge.

Within this setting, the military manager who directs an evaluation of the total system or the subsystem is likely to accept more wholeheartedly the research product when it is expressed in quantitative units that can be related to his goals and missions. The total impact on the operation is the key concern of the military consumer. We believe that human factors research scientists must think in terms of the total mission effectiveness of a system, rather than exclusively in terms of the effective performance of individuals. It is because of the military consumer's end product orientation that systems research and systems development are today enjoying enthusiastic support.

On the surface, the systems output criterion resembles the situational performance criterion, in that both include aspects of the actual job. But development of the systems output criterion requires painstaking experimentation in the laboratory, before taking the criterion into the field, in order to establish quantitative relationships between actual independent variables and various aspects of human performance in the system. In the situation performance measure, subject matter experts are traditionally employed to help assure accuracy of simulation for realism and adequacy of performance coverage. In developing the systems output criterion, operating field personnel are used to help assure adequacy of simulation and coverage, and, equally important, to assist in establishing critical parameters of performance for simulation. Measures of system performance usually involve some clearcut base against which to evaluate performance; for example, accurate and rapid detection and identification of aircraft and tanks.

We think the most exciting and interesting aspect of human performance oriented systems research lie in the near future. There are possibilities for research in the broader areas of social, governmental, environmental regions--to include man-machine systems--in relation to each other and the system and subsystem output. The basic framework of human performance systems research reflects a philosophy of integrated research effort (Uhlaner, 1975). Such a framework is in keeping with the present day direction of systems psychology (DeGreene, 1971), with greater emphasis on application of psychological principles. This framework provides a particular segment of society, in this case the Army, with usable results for the development of effective human performance systems.

REFERENCES

Banks, J.H., Hardy, G.D., Scott, T.D., Kress, G., & Word, L.E. REALTRAIN validation for rifle squads: Mission accomplishment. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Research Report 1192, 1977.

Boycan, G.G., & Rose, A.M. An analytic approach to estimating the generalizability of tank crew performance objectives. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Research Memorandum 77-21, 1977.

Browning, R.C., Campbell, J.T., Birnbaum, A.H., Campbell, Y.A., Fold, G.H., & Haggerty, H.R. A comparison of the validity of officer ratings rendered by hard and easy raters. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Research Report 908, 1952. (a)

Browning, R.C., Campbell, J.T., Birnbaum, A.H., Campbell, Y.A., Fold, G.H., & Haggerty, H.R. *A study of officer rating methodology. IX. Validity of ratings by hard and easy raters*. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Research Report 908, 1952. (b)

Chesler, D.J., Brogden, H.E., Brown, E., & Katz, A. *A study of ratings obtained from raters with aptitude area scores below 90*. Alexandria Va: Army Research Institute for the Behavioral and Social Sciences, Research Report 933, 1952.

Cohen, S.L., & Turney, J.R. *Results of an organizational diagnostic survey of an Army field facility work environment*. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Paper 272, 1976. (AD A020934)

Cronbach, L.J., & Gleser, G.C. *Psychological tests and personnel decisions* (2nd ed.). Urbana: Univ. Illinois Press, 1965.

DeGreene, K.B. (Ed.) *Systems psychology*. McGraw-Hill, 1971.

Downey, R.G. *Utilization of associate nominations in the U.S. Army training environment*: Ranger course. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Research Problem Review 76-8, 1976.

Drucker, A.J. Predicting leadership ratings in the U.S. Army. *Education and Psychological Measurement*, 17, 2, 1957.

Duffy, P.J. *Development of a performance appraisal method based on the duty module concept*. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Paper 273, 1976. (AD A030702)

Farrell, J.P. *Measurement criterion in the assessment of helicopter pilot performance. In Proceedings, Aircrew Performance in the Army Aviation conference at U.S. Army Aviation Center*, Fort Rucker AL: 27-29 November 1973. (AD A001539)

Gaylord, R.H. *Conceptual consistency and criterion equivalence: A dual approach to criterion analysis*. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Research Note 17, 1953.

Haggerty, H.R. *Personnel research for the U.S. Military Academy*. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Research Report 1077, 1953. (AD 21600)

Johnson, C.D. The reliability of averaged ratings with varying numbers of raters. Paper presented at the annual meeting of the American Psychological Association, 1956.

Karcher, E.K., Jr., Campbell, J.T., Falk, G.H., & Haggerty, H.R. A study of officer rating methodology. VI. Independence of criterion measures from predictor variables. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Research Report 905, 1952.

Karcher, E.K., Jr., Winer, B.J., Falk, G.H., & Haggerty, H.R. A study of officer rating methodology. V. Validity and reliability of ratings by single raters and multiple raters. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Research Report 904, 1952.

Macready, G.B., Steinheiser, F.H., Jr., Epstein, K.I., & Mirabella, A. Methods and models for criterion-referenced testing. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Paper, in press.

Maier, M.H. Development and evaluation of a new ACB and aptitude area system. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Research Note 239, 1972. (AD 751 761)

Medland, F.G., & Olans, J.L. Peer rating stability in changing groups. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Research Note 142, 1964. (AD 601 972)

Mohr, E.S. Acceptability of associate ratings at branch schools. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Paper 268, 1975. (AD A017437)

Parrish, J.A., & Drucker, A.J. Personnel research for officer candidate school. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Research Report 1107, 1957. (AD 15507)

Root, R.T., Epstein, K.I., Steinheiser, F.H., Hayes, J.F., Wood, S.E., Sulzen, R.H., Burgess, G.G., Mirabella, A., Erwin, D.E., & Johnson, E. III. Initial validation of REALTRAIN with Army combat units in Europe. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Research Report 1191, 1976. (AD A034610)

Seeley, L.C., & King, S.H. Effects of mandatory showing of ratings to rated officers: Phase II - First Lieutenants. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Research Memorandum 56-21, 1956.

Shriver, E.L., Griffin, G.R., Jones, D.R., Word, L.E., Root, R.T., & Hayes, J.F. REALTRAIN: A new method for tactical training of small units. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Report S-4, 1975. (AD A024030)

Turney, J.R., & Cohen, S.L. The development of a Work Environment Questionnaire for the identification of organizational problem areas in specific Army work settings. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Paper 275, 1976. (AD A038241)

Uhlaner, J.E. Systems research - opportunity and challenge for the measurement research psychologist. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Research Note 108, 1960.

Uhlaner, J.E., & Drucker, A.J. Criterion for human performance research. Human Factors, 1964, 6, 265-278.

Uhlaner, J.E. Human performance jobs, and systems psychology - the system measurement bed. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Report S-2, 1970.

Uhlaner, J.E. Management leadership in system measurement beds. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Report S-3, 1975. (AD A021888)

Wherry, R.J., Ross, P.F., & Wolins, L. Analysis of methods for determining equivalence of criteria. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Research Note 30, 1954.

Willemin, L., Rosenberg, N., & White, R. Validation of potential combat predictors: ZI results for infantry. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Research Note 76, 1957.

Zeidner, J., Harper, B.P., & Karcher, E.K. Reconstruction of the aptitude areas. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, Technical Research Report 1095, 1956.

FOOTNOTES

Extraneous remarks by Mr. Camm

1. Originally, I had two charts, 1945 to 1955 and 1955 to 1976, and they show this trend. The nature of the data is pretty rough. These categories aren't mutually exclusive, so I simply collapsed them into one table.

2. We are trying to get to our construct validity, and this seems to be one way that we can do it.

3. The references here range from 1957 to 1977. The external criteria here in combat situations is combat training like AIT and ratings by platoon sergeants and commanders in places like Korea and Vietnam. Leadership and West Point were based on the same thing; on West Point graduates, how well they performed in West Point, how they were rated by their peers, how well they did after they got out into the field (quite a bit later). The Ranger study is our most recent and has to do with ranger training, peer ratings during ranger training, and how well they performed in Vietnam based on the rating of their immediate commander, usually. We had one more that had to do with the peer ratings of selection for General--but we really haven't put that one together yet. We don't know whether the colonels are rating other colonels on the basis of knowledge of their performance and how good a colonel they are, or whether they know the system well enough to be able to predict who will be promoted to General. We have a lot of problems with peer ratings. They are not very well accepted at this time by people in the Army, and there are a number of complicated reasons for this.

4. There have been several Court rulings that have aided this popularity.

5. REALTRAIN is extremely popular with the troops. We're using it in Europe with great success.

6. TOW is a Targeted Optical Wireless Anti-Tank Weapon.

7. We only have two regiments rigged up like this. As you can imagine, it's a little bulky and inconvenient, but it seems to work quite well.

8. An individual soldier can accomplish the required objective, but he may not accomplish it in the right way, so you have to have somebody out there to watch him.

9. Organizational Effectiveness in the Army has been so successful up to this point that we are developing Organizational Effectiveness Research teams in the Army and sending them to various areas.

10. We're trying to avoid a Hawthorne effect.

11. There's an evaluator in the helicopter itself, and then there's another helicopter that flies about 1,000 feet above with another evaluator. So it's evaluated by at least two people in flight.

12. A lot of missions that the UH-1 pilots perform are at twilight or dawn. One of the problems has to do with the point in darkness that a pilot can successfully perform NOE missions. It was thought that experienced helicopter pilots would have no difficulty with NOE flying. This turned out not to be the case. Pilots trained in NOE could perform; pilots not so trained had difficulty.

13. Q: Is there any device for carrying REALTRAIN kinds of data back as far as the selection level or is it only a training evaluation procedure and it stops there?

A: At the moment, it is a training evaluation procedure, but they are working on carrying it back to at least a selection level. But at the moment it's strictly a training evaluation procedure.

Q: How is your skill qualification test coming, and what do you estimate to be the cost per year of operationalizing it and managing it?

A: The skill qualification testing is coming along great. We'll have the SQT's in place in about a year and a half or two years. I have not even the foggiest idea of what the cost is.

VIII

NAVY EFFORTS IN CRITERION DEVELOPMENT FOR JOB PERFORMANCE EVALUATION

Frederick A. Muckler
Navy Personnel Research and Development Center

Introduction

One nice thing about discussing the area of criterion development for job performance evaluation in the Navy is the multitude of available examples. Indeed, all of our systems applications and our R&D programs are, without exception, infested by the criterion problem. Thus, my charge--which is an "Overview of US Navy Efforts in the Criterion Area"--is in one sense a simple one. I can state categorically that where we have a human behavior measurement program we have a criterion problem.[1]

Further, in general, we adopt one of three approaches to the criterion problem. First, we often ignore it and hope that somehow the solution will appear as a natural result of doing the work. Second, we often agonize over it. The question most often heard here is: "What does all this mean?" Third, we may attempt to solve the problem scientifically; this is the "sound methodology" approach which assumes that good methods will extract acceptable criteria. None of these approaches, of course, tend to work very well,[2] even where in many cases we will alternate between all three.

The basic problem, it seems to me, is that we persist in demanding meaning from our measurement. We want to be able to know what our job performance measures add up to; we want to evaluate them. If we only did not have to do that--if we could only be satisfied with the data points alone--the criterion problem would disappear. Indeed, some of us adopt just that technique. We collect the data, publish the report, and leave the meaning to somebody else. Unfortunately, we have all found that when others interpret our data the consistent result is misinterpretation and misuse.

From a host of possible topics of concern to Navy research, I would like to concentrate today on three areas. First, we are concerned with methods of generating criterion sets; I shall be concerned with four tools and the problem of "criteria of criteria." Second, I have selected six specific technical problem topics with the criterion development area. And third, I would like to mention seven applications examples where the criterion problem remains unresolved.

52

So far as I can see, while the areas reviewed and the examples cited are Navy-specific, all of them represent problems in criterion development for any context of human performance evaluation. I do not see that the Navy has any unique problems in this area. Rather, they are problems shared by all and, sadly, they are problems which have had a persistent history in industrial and organizational psychology (Gilmer, 1971; Landy & Trumbo, 1976; Smith, 1976; Thorndike, 1949).

Generating Criterion Sets

[3]With respect to the first area--that of generating criterion sets--I will assume that we have available some quantity of raw job performance data: a lot or a little, subjective or objective, complete or incomplete. Given those data, the question now is: "How do we evaluate it?" Or "What does it mean?"

Technically, it seems very important--to me at least--to repeat again and again one fundamental point: the measures of job performance and the criteria on those measures are not the same thing. Criterion "measures" are in fact above and beyond performance "measures." Performance "measures" are neither good nor bad; criterion measures make them so.[4]

[5]Smith (1976) has recently commented: "The first requirement of a criterion is that it be relevant--to some important goal of the individual, the organization, or society."[6] If one accepts this requirement, it seems apparent that criterion sets are transforms on the job performance measure sets. These transforms must relate to domains far beyond specific job performance per se.

So, our problem here is the methods by which we generate criterion sets which in fact will provide judgement, if you will, to some other context. I would like to distinguish four general methods, all of which can be seen in current Navy research and development.

(1) *"Traditional" sets*. I doubt if there is any context in which we work with job performance measurement where there is not already a "tradition" of past criterion sets. One of the major emphases of many current Navy R&D studies is "productivity" (Muckler, 1976). We are concerned with the lack of it in Navy task performance, and we are much concerned with methods of enhancing it. The criterion may be simply stated as: *More is better*. Whatever the individual does, he or she should do more of it in the same unit of time.[7]

But in most cases, "more is not better." I am reminded of a productivity enhancement program in a cigar manufacturing plant where individual cigar output per day was increased from 3,000 to 6,000 per day by using all of our bag of tricks in self-pacing, participative management, work incentives, and so forth. Unfortunately, the sales manager returned to the plant and informed management that the plant aggregate based on 3,000 per day per worker was all the market could bear. The end result of 6,000 per day was a lot of cigars stored in the warehouse, so more is not necessarily better.

A second example concerns the productivity of our training systems. Navy programs are no exception here to the demands now being placed on all training systems everywhere: We are told that we must have more and better training for the dollar. With respect to more training, certain traditional measures suggest themselves immediately: (1) number of students produced, (2) staff/student ratio, or (3) attrition rate. We must maximize the first and minimize the second and third. Unfortunately, none of these seemingly useful traditional measures has clear criterial interpretation. How many students we produce, for example, must be tempered by how many students we place in jobs.[8] Further, to state that a training activity has attrition rates of 0%, or 50%, is meaningless without reference to other criteria. I assume that should we achieve 0% attrition we would then be accused of making training too "easy."

The difficulty with traditional measures is that while they may be incomplete, ambiguous, or even incorrect to us, they are often most "relevant" to others. In job performance, for example, it is natural that managers should ask for more productivity; they are most often judged on the basis of that single, "ultimate" criterion. We must, I think, at least be sympathetic where "simple" criterion measures are commonly used.[9]

(2) "Theoretical sets". How delightful it would be if we had formal quantitative models where the criterial transforms would be clearly and mathematically specified. We would know what they are and how they are computed. Considering the sheer amount of past work in job performance evaluation covering surely thousands of research publications, it may seem strange that we do not have more formal theory. In some few selected cases such theory is available, but even here the issue is not simple.

It was my pleasure for some years to work in an area where the relationship between individual job performance and system performance could be mathematically stated with great precision. This was the area of optimal control theory. Given the statement of the system state spaces and the allowable system processes, it is possible to define mathematically optimal paths. But even here the judgmental process was essential. It turns out that there is no one optimal path for any system. It depends on what you want. And what you want depends on judgments that have nothing to do with the measures or the mathematics.[10]

To my knowledge, we have no R&D programs working on developing quantitative theoretical models that will relate our job performance measures to our criterial sets. The closest thing to it has been connected with the computational problem of dealing with very large numbers of predictor and criterion variables. The past decade has brought us both the mathematics and the computer capability to deal simultaneously with very large N-dimensional measure sets. At the present time, we have a program based on complex polynomial regression equations using mini-computer technology specifically designed to deal with job performance measures.

But while these techniques will allow us to handle large quantities and kinds of job performance measures, they are not "theory" in the sense I am using it here. They will allow us to process coherently large amounts of job performance data, but they will not tell us what is good and what is bad.

(3) Empirical methods. To me, one of the most interesting developments over the past decade has been the development of empirical methods of deriving both criterion measures and the weights that should be assigned to those measures. It seems particularly appropriate here that mention be made of the work of Ray Christal and the JAN procedure (1968) and synthetic criterion methods (Mullins, 1970). With this technique, and others like it, the logic seems clear: If criterion sets require expert judgment, then let us systematically and empirically investigate the experts.

It would appear that the most popular technique at present with Navy programs is Delphi, the procedure normally associated with Dalkey and Helmer (1963), and the Rand Corporation. For some reason, Delphi has become extremely popular in Navy programs. Recently, I have seen Delphi used in such situations as decision making, unit performance measurement, training, tactical field exercises, and the like (Sander, 1975; Larson & Sander, 1975). There is certainly something very satisfying in a systematic way of collecting expert opinion and using this to deliver criterion sets. The results always seem to me to be very interesting.[11]

But at the risk of seeming simple-minded or, worse, anti-empirical, something always bothers me about these studies. I find myself constantly asking the question: "Is this really true?" Or, perhaps better, "What is the probability that even a large group of experts can come to the wrong conclusions no matter how carefully their judgments are collected?" Or, another question: "Do 'subject matter experts' really know what the problem is?" In short, just how much confidence can I place in the validity and completeness of criterion sets generated by experts?

A case in point. I suspect that if I were to use Delphi on industrial managers, the result would be that the most important single criterion is to maximize profits. Yet studies by Stagner and many others have shown very clearly that in fact they do not behave that way. They simply do not behave as managers to maximize profits. What they say and what they do are not necessarily the same thing. Delphi may give me what they say, but is it what they do?[12]

(4) Criteria for criteria. Last, I would like to turn to criteria for our criteria. Those of us trained in traditional psychology, I hope, surely cannot ever forget validity and reliability as criteria for our criteria.[13] But the literature of the past few years seems to me to raise the question of "completeness." Validity and reliability are surely necessary, but they seem to be not sufficient.

Let me quote again from Smith: "The first requirement of a criterion is that it be relevant--to some important goal of the individual, the organization, or society." Somehow I feel that our traditional methods of demonstrating validity and reliability will be insufficient to satisfy that requirement.

Fortunately, the American Management Association Management Handbook (Moore, 1970) provides a set of criteria about criteria from the management point of view. There are eight of these, and I would like to apply them to the problem of job performance evaluation.

(1) Suitability. Are the measures relevant, and do they support the purpose and mission of the organization?

(2) Feasibility. Are the measures theoretically attainable within the organization?

(3) Acceptability. Will the management accept the measures and provide the resources to collect the measures?

(4) Value. Are these measures the best buy for the money?

(5) Achievability. Can, in fact, the measures be collected?

(6) Measurability.[14] Can the measures be quantified in terms of quality, quantity, time, and cost?

(7) Adaptability and Flexibility. Can we change the measures to reflect changing organizational environments and management needs?

(8) Commitment. Does everybody in the organization want to do it?[15]

This, then, is one management view about the evaluation of our job performance measurement. Frankly, considering how difficult it has been for us just to get marginal validity and reliability for our measures, these additional eight requirements seem rather overwhelming.

Some Current Technical Problems

Let me now turn to the second topic area. I have selected some six issues that bother us. The list is by no means exhaustive, but there are problems, as I look across Navy programs, that I really see looming very large.

(1) Data acquisition. First, the problem of collecting data. It seems to me that with respect to job performance evaluation, we are routinely collecting more and more data points. For several reasons, it seems a great deal easier to collect more and more data. Indeed, it seems to be expected.[16]

In a current study we are collecting data on over 50 measurement dimensions for the job performance evaluation of sonar technicians. Included are cognitive, vigilance, noncognitive, biographical, perceptual, biochemical, standard test, and peer rating measures. The principle seems to be: If it moves, measure it.[17]

(2) Data processing. We feel free to measure more and more things because we now have available (theoretically) enormous data processing capability. To be sure, thanks to the computer, we can now do data processing tasks that simply could not have been done manually a decade ago.

This is certainly true for our studies in job performance evaluation.[18] We can use standardized scenarios to measure job performance through computer training modes. And, as another study has shown, some minority group members perform better than they do in the traditional evaluation situation.

(3) Cost effective criteria. But all of this is not at small cost. It seems reasonable (indeed, essential) that we ask if all these additional data points and these computers are cost-effective. I do not know. I do know the data acquisition and processing techniques we have been exploring are far more expensive than "traditional" job performance evaluation methods.

In some cases, we are introducing job performance evaluation where there has been none before. The cost comparison is particularly unfortunate: zero versus N-thousands of dollars. The expression of effectiveness for these costs is not certain. In one specific case,[19] we were able to disclose certain critical skill deficiencies and institute remedial training to eliminate those deficiencies. Was it worth it? That is difficult to say.

(4) On-the-job validation. On-the-job validation of job performance evaluation has always been difficult. On the one hand, we appear to be getting much better access to the operational environment. We are doing better aboard ship, and where that is not possible, we are bringing very sophisticated measurement vans dockside to the ships.

On the other hand, there remains a large core of job performance measures that we cannot validate without World War III. One increasing trend here is the use of full scale simulation of the mission as the validation device. While I see no alternative at the present time, one is left with the doubt that performance in the simulator may or may not predict performance in combat.

(5) Simple versus multiple criteria. Next, no one likes simple measures more than I do. Yet I do not see how we can ever expect to get simple criteria for a process as complex as human job performance. Looking only at the task itself and the performance associated with it, I have yet to see a "simple" task or "simple" performance. I sincerely hope I am wrong.

I cannot pass this subject by without commenting on the Holy Grail of job performance evaluation: The Ultimate Criterion. In the literature, and certainly in practice, we continue to hope for that single, final, criterion that will express everything--whatever that may be (Thorndike, 1949). But it seems to me that researchers at least have abandoned that search. Every current study of which I am aware assumes the need for multiple criteria.[20]

(6) Measurement versus evaluation. I am still concerned, however, with what appears to be a continuing confusion between job performance measurement and the evaluation of that measurement. We appear to be in a minor phase of, as just noted, radical expansions in the quantities of data we collect. I would predict that this phase will begin to change and that we will, in the future, be collecting less data. We are, I hope, going to become more discriminating in getting that data relevant to interpretation and use.

Some Criterion Application Areas

Let me now turn to my last area which is some of the specific application areas in which Navy research and development is under way. In each of these cases, it appears to me increasingly that the question is being asked: "What do you want to know?" before we decide what job performance measure sets we should collect. Depending upon the use of what will be made of the data, it seems clear to me that differential job performance measure sets may be selected. Or, to put it another way, in each of these cases job performance evaluation is essential, but the measure sets may differ depending upon the application. Incidentally, I have yet to be able to convince many of my colleagues that this might be true. So let me offer it to you as a possible hypothesis.

(1) Individual job performance evaluation. I have made several mentions about individual job performance evaluation. Let me summarize as follows: We are taking much more complete measure sets, we are doing much better in job performance evaluation in operational environments, but we have yet to demonstrate convincingly (at least to me) that we are cost-effective.

(2) Unit (team) performance evaluation.[21] Increasingly, our efforts are turning (or perhaps returning) to the importance of unit (team) performance measurement. A very positive sign to me is the renewed attempt to measure both process and outcome of team performance measurement. For some time it seemed to me that we avoided outcome measurement because it was so difficult. For example, studies of communication systems stressed all sorts of internal process measures such as frequency of interaction and so forth, but I never knew what happened to the messages. In this case, the Delphi technique appears to be useful in deriving unit performance effectiveness measures (Larson & Sander, 1975).

(3) Personnel subsystem readiness. Many of our users are not satisfied with evaluations of individual job performance. We have been getting increasing demands for some expression of the state of the entire personnel subsystem (Borman & Dunnette, 1974). We are asked, for example, "What is the personnel readiness of this ship?" In short, what is the aggregate of all the people on the ship? I would not pretend that we have an answer to that question, but we are trying to see what we can do with the question. I, myself, am not yet convinced intellectually that it is a meaningful question, but emotionally and intuitively, I find it very attractive.

(4) Personnel/system operational readiness. To move up one level of complexity, we are increasingly being asked to contribute to some representation of total system operational readiness. In terms of operational readiness, for example, what does it mean when the ship is 95% manned? Or, what does it mean if the personnel in a given rate are only 75% job proficient? I would not pretend that we know how to answer these questions precisely, but we are being asked once again. At the present time, the method primarily in use is through total system simulation models performance. I hasten to add this is modeling simulation and not physical simulation.

(5) Selection, training, and organizational development. In the areas of selection, training, and organizational development, I find a number of what are to me encouraging trends. For one, the performance measurement seems to me to be getting far more precise and hence of much greater diagnostic value (Campbell et al., 1974). This is praticularly true in training. Job-referenced performance measurement seems to me to be looking much closer at the microstructure of job deficiencies. This is not for the sake of measurement, but rather so that remedial training can be closely tailored to the individual's training needs. In organizational development, it seems to me that performance measurement is becoming far less global and vague and far more sensitive to the actual events that occur--complex though they may be.

(6) Productivity and accountability. I have made previous mention of the problem of productivity. In this case we are being asked to supply job performance measurement that will serve as the basis for

productivity enhancement and individual team and organizational accountability. I, for one, am glad that we are being asked. We remember, I hope, how job performance measures have been misused in the past for these purposes. If we only stop people from repeating past mistakes, our services will be of value.

(7) Evaluation of R&D personnel. To end on a threatening note, we currently have underway studies on job performance evaluation of R&D personnel. In a program called SHORTSTAMPS (or Shore Requirements, Standards, and Manpower Planning System), the Navy is attempting to perform job performance evaluations on all Navy shore personnel with the objective of better staffing standards and use of manpower. Since R&D personnel are a part of the Navy's shore manpower requirements, it seemed reasonable to management that R&D personnel should be included. I assure you that we argued vigorously against this assumption, but to no avail. Since we lost, we have decided to help them.

I am reminded of a statement once made to me by a manager: "Somebody is going to have to guess, and your guess is better than ours." I think he was right.

REFERENCES

Borman, W.C., & Dunnette, M.D. Selection of components to comprise a Naval Personnel Status Index (NPSI) and a strategy for investigating their relative importance. Minneapolis: Personnel Decisions, 1974.

Campbell, J.P., Bownas, D.A., Peterson, N.G., & Dunnette, M.D. The measurement of organizational effectiveness: A review of relevant research and opinion. San Diego CA: Navy Personnel R&D Center Tech. Rep. 75-1, July 1974.

Christal, R.E. JAN: A technique for analyzing group judgment. *Journal of Experimental Education*, 1968, *36*, 24-27.

Dalkey, N., & Helmer, O. An experimental application of the Delphi method to the use of experts. *Management Science*, 1963, *9*, 458-467.

Gilmer, B. von H. *Industrial and organizational psychology*. New York: McGraw-Hill, 1971. Pages 353-360.

Landy, F.J., & Trumbo, D.A. *Psychology of work behavior*. Homewood IL: The Dorsey Press, 1976. Pages 88-130.

Larson, O.A., & Sander, S.I. Development of unit performance effectiveness measures using Delphi procedures. San Diego CA: Navy Personnel R&D Center Tech. Rep. 76-12, September 1975.

Moore, R.F. (Ed.) *AMA Management Handbook*. New York: American Management Association, 1970. Pages 1.31 to 1.33.

Muckler, F.A. Productivity evaluation and measurement. Paper presented at The First International Learning Technology Congress and Exposition, Washington DC, 21 July 1976.

Mullins, C.J. *Estimation of validity in the absence of a criterion* (AFHRL-TR-70-36). Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, October 1970.

Sander, S.I. Delphi: Characteristics and applications. San Diego CA: Navy Personnel R&D Center Tech. Note 76-2, October 1975.

Smith, P.C. Behavior, results, and organizational effectiveness: The problem of criteria. In M.D. Dunnette (Ed.) *Handbook of industrial and organizational psychology*. Chicago: Rand McNally, 1976. Chapter 17: Pages 745-775.

Thorndike, R.L. *Personnel selection*. New York: John Wiley, 1949. Page 121.

FOOTNOTES

Extraneous remarks by Dr. Muckler

1. On the negative side, in looking back over the past 4 or 5 years of Navy research, I find, to my dismay, that at least for the last 2 years there has been no program, principally or primarily, concerned with the criterion problem, per se. As a matter of fact, those programs which I would classify in that area sort of began to die about 1974. One particular example was Dr. Campbell's work for us in Measurement of Organizational Effectiveness which was a super job. The problem is not that we have not proposed such programs, but rather that we have not been able to sell them, and I think that Col Fulgham's distinction this morning was a very good one. The programs we have proposed have been considered to be, "Yes, they would be useful but not usable." And so it's been a real problem for us to convince our own people that it would be useful to do work in criterion development, despite the fact that again there is not a program we have which is not infested in some way with a criterion problem. I'm also concerned as I walk around and talk to all our researchers and I talk to the people who do our research--all of them, of course, presumably very competent psychologists--at how many of our researchers do not recognize the criterion problem exists. And I think if you think back, if you were very careful to avoid a course in industrial psychology or courses in psychometrics that you could pass through the PhD program without ever having come in contact with the criterion problem. And so for those of us who live and die by this problem and who are fascinated by and concerned by it, it is a little alarming, I think, to see a researcher in fact embedded in an enormous criterion problem without any awareness whatsoever that that problem exists. If I look across our programs and see what our people do with the criterion problem I find one of three approaches being used and sometimes all three.

2. They tend to work even less if you try them after the program has started.

3. It is my unfortunate tendency in discussing research, particularly with our research workers, to ask many questions about their research. One question that I continue to ask along the line is, "Why are you measuring that?" And I've discovered that I'd better ask that question very carefully because frequently I get a response which implies, "What the h--- are you talking about?" Or, I frequently get a hostile response, "What's wrong with that?" And, of course, the answer usually is, "Lots." But I generally stop asking at that point.

4. I think this is more than just a semantic point. It seems to me that an awful lot of the confusion in existing literature and even

among ourselves would be not perhaps resolved but would be clarified if we were very careful to distinguish two levels of description. Unfortunately, we've sort of settled into this multiple regression approach and we call these predictor variables. That's all right--of course most of them aren't--but that's all right if we call them that. But we have gotten into the habit of calling these criterion variables, and maybe someone gave some of those definitions this morning--that's okay, there's nothing wrong with that--but it seems to me that it would clear the air a little bit in a lot of cases if we would separate that into two levels of description. And what are the output measures, or what is it, what's happening? I wish we would go back to the normal use of the word "criterion." I wish we would realize that, in fact, when we're talking about criterion measurement, as we will, we are talking about the standards or values on the output measures; that in fact, our criterion measure is our transformance on the output measure; and furthermore that an output measure, a behavioral measure, does not contain necessarily within itself any meaning of good or bad. It seems to me very frequently we take a measure and we assume without being explicit about it what's acceptable and what is not. It seems to me if we were very clearly distinguishing between these two levels of description, a lot of the confusion would clear up. If I might take for an example "errors." It was my misfortune--no, I shouldn't say that--I happened to be present by accident with the start of the zero defects program. It was really, truly accidental. And what started out as a very nice idea--the goal of zero defects--somehow got transformed into the requirement for zero errors. And because we are vague and not too explicit about this, people began to say, "Gee, we've got to have zero errors." I don't know of any human activity where you're ever going to have zero errors, and, what was a reasonable goal is an unreasonable requirement. But it seems to me that frequently when we take error measures we automatically assume that zero is good and I would argue to you that that is not necessarily so. And when we looked at the errors that existed, then the first question was, "How do you reduce the errors?" And, obviously there are many ways of doing this, but associated with that is some cost function. And, in many cases, we've found that there was no question that one could reduce the errors, but as we began a minimization function on the errors, that the cost of so doing increased very erratically. So we began to get that sort of thing. I would argue to you that the error is the output measure, the criterion measure is really this cost-function. And then the question becomes much different when one is looking at it this way; much different about this sort of desire of having zero errors. In fact, what one then does is make a judgment and say, "I'll accept that level of errors as being acceptable within my system," (right off the bat that makes you have to define it--define what level of error is tolerable) "and for that I am willing to pay that much." I don't want to belabor this--I will, of course--but I really think it would help an awful lot if we did make this distinction. I really think it would help a great deal. And particularly now where our measure sets are being imposed upon by many other than our traditional criteria (some of which I will get to).

5. Patricia Smith, in her article (which Major Sellman mentioned)--May I call this a mini-stop now for a promotional plug on the Dunnette handbook which I think is one of the finest things that's ever appeared for our field. I wish it had been a little lighter and, of course, a little cheaper, but that's the way it goes, isn't it? That's the cost-functional on it.

6. Relevance to the individual, to the organization, and to the society. I don't see where any of that is contained in, say, an error measurement. Indeed, it is a separate transform on those error measurements. So our problem here which some of us, at least in the Navy, are much concerned about, is how do we develop all these measure sets. How do we develop the output measures, but more than that, how do we develop the criterion transforms on those measures. And the answer to that is, "Very badly." There are four ways that I see that we do this sort of thing. The first, trying to be as kind as I possibly can, is the traditional way.

7. This reflects the Navy's almost frantic interest in productivity. Everybody is concerned about the productivity problem, but I think we have gone beyond concern into hysteria--with good cause, I might comment. We have some rather large organizations in the Navy that are setting new records for non-productivity. As a matter of fact, we wouldn't mind that very much if they stopped making trouble too. Sort of the optimal combination. I have a great deal of trouble explaining to people that they might consider the possibility that more is not better. It does not necessarily imply that because we have more output that this is better. It seems again that there's a confusion between the output description and the criterion measurement judgment.

8. We've been having a very interesting problem in some of the individualized self-paced training programs that we have done. They have been extraordinarily effective. They have, in fact, produced very high quality students in the sense of the very excellent measures of their proficiency, but they have wreaked havoc with our logistic system. One student comes out in 3 weeks, and the next student comes out in more or fewer weeks. The manpower allocation system has just been thoroughly and totally confused. Another thing that Lee mentioned this morning - the goal there is 100% proficiency, and by God we get them there and then we no longer have any variance on them. In one particular case in which I'd better leave out names since it involves Admiral Rickover, there is a concern about the fact that we give

them a set of students where they're trained to 100% proficiency and they say, "Well, how can we discriminate between them?" And we say, "You don't have to." And then, "No, I don't believe that." So here we've got a measure where we get everybody 100% proficient and, in fact, it's not acceptable to the operational people. We are under a great deal of pressure to reduce attrition rates. There again, the question is, "What's an acceptable attrition rate for anything?" If you don't really carefully distinguish between these two things you sort of automatically assume zero attrition is what you want. I would argue not so. Zero attrition, 25% attrition, 50% attrition, those numbers in themselves have no evaluation--they're neither good nor bad. It really depends on what your system wants to achieve.

9. It would be awfully nice, I think, if we had the kind of formal quantitative mathematical theory which would, in fact, define and set both our measures and the transforms on them. In most cases we do not have this. And in those cases where I have worked where we do have this, even that hasn't solved the problem.

10. So you started off this whole modeling business by saying, "What is it in your subjective judgment that you want to have?" Once having made that clear, then we can crank the whole model out and we can tell you how to go the best path based on that objective. I don't think that in my life time I'm going to see that kind of theoretical development in our area and, in lieu of that, I suppose we ought to just muddle through--and I'm sure we will. I think it might be worthwhile to comment here just a little bit, if I might. At a point in my career I had to work a great deal with mathematicians working in modern optical theory and the mathematics are just super. You can spend a whole week looking at an equation. It's the best of all possible partial differential equation work and if you get your jollies that way, that's where you get them. I discovered to my surprise that many of those models don't predict anything. No, I take that back. They predict a lot of things which aren't true. In my experience in several areas of physical theory--you know that hard stuff we always talk about--a lot of their models are not correct. They simply are not valid; and it doesn't seem to bother them. In acoustical theory, I commonly saw the pattern where everybody set up the equations, there was a big computer study, predictions were made, and then they set up a simulation that fixed it the way they wanted it to be anyway. It's interesting that psychologists, it seems to me, have been extraordinarily concerned about what we're doing, and the quality of what we're doing, and the meaning of what we're doing, and I think that's very, very good. On the other hand, it seems to me that very frequently we get upset because our problems are so complicated that it seems to us it's all unsolvable. As far as I'm concerned, having worked in many other theoretical

areas, I think psychology's in pretty good shape. I wish we wouldn't cry so much about it, however.

11. In the Delphi application to tactical field exercises, the set of measurable criterion dimensions was, I thought, really quite sophisticated.

12. I shouldn't tell this story because it's not a very nice one. You recall that these techniques have one basic technique that was used. And that technique was that we want to collect these data from the experts independently and anonymously cause we know what happens when you put them all together in one room. A very recent study was done which I did not know about until after it was done in the Navy. They didn't have time to do that and they had them all together so they sat down and they did it in one room, and there was, in fact, a hierarchical rank system operating. I'm also reminded of a study I did some years ago in flight test of an instrument. We had 12 flight test pilots--from a service I will leave unnamed--evaluate that instrument. They sat down as a committee to evaluate the instrument and they said, "How many are in favor of this instrument?" The first vote was 11 to 1. The one vote was, unfortunately, the commanding officer, and he said, "We will now have a second vote." The second vote was 0 to 12. I'm astonished; I thought everybody knew about that sort of problem.

13. Obviously, we're very much concerned with this problem for any measures that we take. Beyond that, we talk about other things like contamination and deficiency. I prefer to think of deficiency in terms of the completeness of the measure sets. How complete is your measure set to describe the phenenoma that you're dealing with--but that's another problem. I'd like to talk a little more about this because based on Smith's definition where the criteria must be relevant to the individual, or the organization, or the society, we might ask some questions about what kind of criteria could you get that would define that relevance. How can we say for example, "How would the organization view our criterion measurement?" "What sort of criteria would they put on our criterion?" Needless to say, that literature is not a very large one, and it's sort of like - this is a good thing to do, but nobody's been explicit about what these criteria might be.

14. Are you going to give me measures that I can do something with? And it was interesting in this particular management handbook that concern was with both measures of now and also measures of the time history. I thought that was extremely interesting and extremely sophisticated. If we recall some of our own literature here (Dr. Camm has contributed about the dynamic nature of criteria), it seems to me they did not acknowledge you but it

seemed to me like it was awfully nice there was concern about an understanding of the fact that criteria are not eternally stable.

15. Of course the answer is no, no matter what organization you have. We're engaged in our annual orgy of performance appraisal at NPRDC, and I suspect if you were to ask about the commitment problem, that we would cease instantaneously to do so. This is not true everywhere. Nobody in particular likes this sort of thing but they do it anyway. These then are how management of the organization might respond, by their criteria to our criteria, a possible set of criteria on criteria.

16. Our users are, frankly, much more sophisticated about this. I think, with many of our users, if we came in and collected one number, one output measure, they would be disappointed. They really expect us to collect large data sets.

17. This is good news and bad news. It's good news because we're collecting a lot of data, and we're collecting it of a magnitude so that we can really do something with it. But of course it's bad news because what it really reflects is we don't know what we're doing. And we're going to make overkill and make sure that we don't miss anything. And so we will have a lot of pseudo predictor variables.

18. In going aboard ship, which is a game we play, we are finding aboard those ships computers. Now they're there for other reasons. And we are finding that they're not being used all the time. And we say "Hey, can we use those computers?" And the answer is yes. So now when we come aboard we bring a terminal and software and we time share with the onboard computers. And we use these in evaluating for many, many purposes. One is, frankly, personnel management. I think you would not be surprised, aboard a carrier with 2,700 people or 3,000 people as the case may be--one, by the way, is never sure how many are aboard--I think you would not be surprised to know that very frequently there is less than optimal allocation of personnel resources. Translated, I remember one propulsion evaluation board on one of our carriers--the PB set up certain standard problems and they expect people to solve them. In this one case, not only could they not solve them but they couldn't find anybody who could. Not because he was not there--the guy was there--they just couldn't find him. Then we're talking about 600 men in the Engineering Division, and just nobody knows where they are. So this is a real problem. By the way I might comment, you don't experiment or test with the devices you take aboard. You plug in with the onboard computers and we find that, really, these are extraordinary opportunities with respect to job performance measurement. So, for example, we can set up

little standard job scenarios, have the folks come down in their off-duty cycle, and we can measure rather directly their job performance, with respect to standard job scenarios. And this is working just beautifully, providing the commanding officer likes it.

19. This was dockside, job performance evaluation in three skilled categories: sonar technician (of course), weather technician, and missile technician. Now these are supposed to be the best guys we've got. They're out there doing their jobs; they've been through all the schools and they've years of experience, and they're supposed to be super. Jerry and his folks went down and tested these people on some very sophisticated job reference tests, and the first thing we found was some rather startling deficiencies in what the very best of our people could do. You know you really don't want a nuclear warhead technician at 70% effectiveness, I think. I'm happy to say immediately that they brought with them remedial training programs tailored specifically to the individual so that the measurement that they got was diagnostic and could, in fact, be used immediately by the people. I'm happy to report from the latest data that this was extraordinarily successful. But it was extraordinarily expensive as well. And so one gets to the point of saying, "You've got a nuclear warhead technician. What is the effectiveness of changing his job proficiency from 70% to 98%?" Well, emotionally, it makes me feel much better. But, is this the kind of data that we can present for cost effectiveness evaluation? I doubt it very much. Well, let me put it this way: It hasn't worked so far.

20. I don't see that this is a problem in practice. It seems to me that in most situations that I'm familiar with, I don't see many people looking for simple criteria. In practice they are really looking for multiple criteria because that's the nature of what you're dealing with. In dealing with mathematicians--it was always an interesting experience for me to take this kind of problem to a mathematician. For two years I was with some of the world-class mathematicians who assured me that no matter how complex the problem was they would find it mathematically tractable. This was, of course, before they saw our problems. And so we started giving seminars to the mathematicians. We started saying, "Okay, here's some of our problems, now what do we do with this mathematically?" I recall one, Ruth Holliman, who's famous for the Holliman filter, who said, "That's too complex." We used to have a little scenario in a special, beautiful mathematical library. You recall Einstein's theory of relativity rested on Riemannian surfaces, which is the theory which had been developed about 20 years before. So he had a model that he needed. We had this little thing that we're going to walk through the mathematical library and a volume would fall on the floor open to Chapter 15, which was

68

the model for our data. This was our theory of divine intervention--and it never happened.

21. With respect to individual job performance evaluation from a summarized sum of the comments, I see much more sophisticated measurement than I've seen, I've seen much more in-depth, on-the-job performance measurement, and frankly something there I like, I see a lot more of "objective" measurement. Mr. Camm noted some of these. We're less and less dependent upon rating methods. I'm really not against rating methods but I sort of like the fact we have much more measurement opportunity in individual job performance situations.

IX

THE CRITERION PROBLEM
AN OVERVIEW OF EVALUATION AND MEASUREMENT RESEARCH IN THE AFHRL TECHNICAL TRAINING DIVISION

Philip J. DeLeo and Brian K. Waters
Technical Training Division
Air Force Human Resources Laboratory
Lowry AFB, Colorado

The Nature of the Criterion Problem in Technical Training

People engaged in training research frequently view the well-known criterion problem from a somewhat different perspective than those who perform selection or classification studies. The typical selection study begins with a careful search for criteria which possess, among other desirable properties, (a) relevance to the ultimate criterion, (b) freedom from contamination, and (c) reliability (Thorndike, 1949). Selection and classification researchers then devise methods of measuring behaviors (i.e., ability or aptitude test performance) which predict the criterion chosen. In contrast, training researchers are likely to accept the criterion objectives of a training course, or unit of instruction, as "givens" and bypass that aspect of the criterion problem completely, choosing instead to concentrate on what is essentially a measurement problem, namely making the mastery or non-mastery decision on specified criterion objectives. Thus, in both the knowledge and performance domains, the criterion problem becomes a question of whether or not mastery of the criterion is the state of nature for a certain individual. Relying on the instructional system development (ISD) process to specify appropriate criterion objectives, training researchers have tended to concentrate their energies on developing methods for measuring whether these criterion objectives have indeed been attained. This strong emphasis on measurement will be seen clearly when we discuss our past efforts, and it continues prominently in our present and planned work.

Having contrasted selection and training approaches to the criterion problem, let us now attempt to show how they are related. Figure I illustrates the linkages between selection, training, and the job in terms of immediate, intermediate, and ultimate criteria.

Most, if not all, Armed Forces selection and classification tests are validated using performance in training as the criterion--for the obvious reasons that training data are easier to obtain, less costly, relatively reliable, etc. But, it is clear that only to the extent that training performance is truly reflective of job performance are

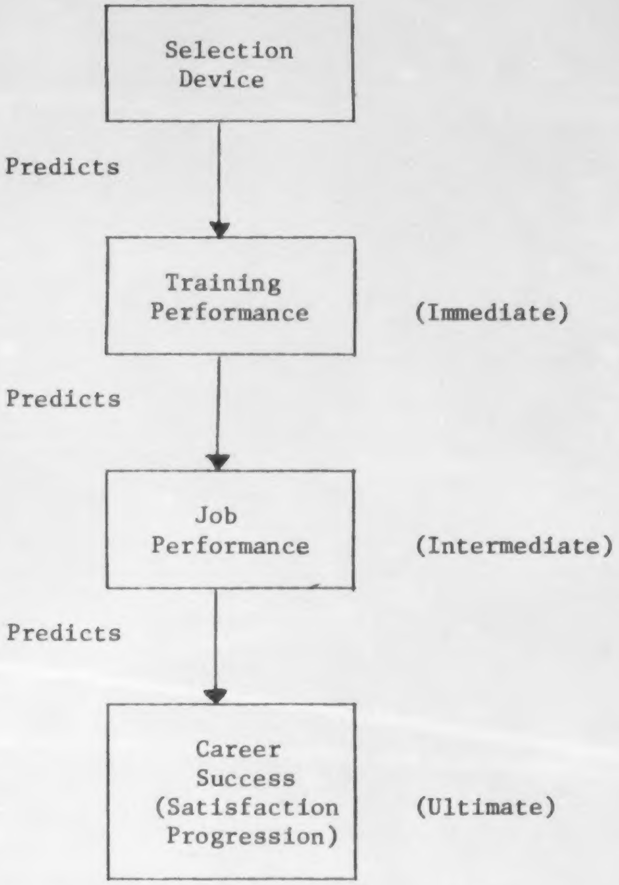

Selection Device
Predicts
Training Performance
(Immediate)
Predicts
Job Performance
(Intermediate)
Predicts
Career Success (Satisfaction Progression)
(Ultimate)

Figure 1. A model of the relationship between selection and the ultimate criterion.

selection studies on safe ground. For the process described in Figure 1 to be valid, it is incumbent on training researchers, therefore, to re-examine a more classical statement of the criterion problem and consider to what extent training performance actually predicts job performance. While accurate measurement of training performance is seen as a necessary condition for total system effectiveness, it is by itself not sufficient. Realizing this, we have increased our emphasis on improving training evaluation (Step 5 of the ISD process), and we will in the future conduct research to improve the methods by which both training requirements and training objectives are developed in Air Force training. (Steps 2 and 3, respectively, of the ISD process.)

To recapitulate, thus far we have asserted that "solving" the criterion problem involves answering essentially two questions:

(a) what behaviors should be observed (measured, tested) and (b) how are these behaviors to be measured effectively (i.e., taking into account reliability of the measuring devices, efficiency, and accuracy)? The decision to observe certain behaviors rather than others involves a content validity approach which is based on defining the job domain in terms of tasks performed. This aspect will be subsequently referred to as the definition aspect of the criterion problem. The question of measurement effectiveness equates to a predictive, or concurrent, validity approach which relates training performance to job performance.

Table 1 provides a complete overview of our measurement/evaluation research work as it relates to these two aspects of the criterion problem. We shall next review these studies in some detail, indicating general trends in our program.

Table 1. The Criterion Problem

	Measurement Aspect	Definition Aspect
Past	o Student Attitudes o Confidence Testing o Advanced Measurement Techniques o Adaptive Testing	o Survey of ATC measurement/ evaluation procedures o Task clustering in field evaluation
Present	o Adaptive Testing Model Development o Symbolic Performance Testing o Criterion Checklist Reliability	o Advanced Field Evaluation System
Future	o Latent Trait Applications o Adaptive Testing Implementation o Criterion Referenced Testing (Mastery/Non-Mastery)	o Requirements Validation o Workshop for Implementation of Advanced Field Evaluation System

Previous Work

Since the Technical Training Division of AFHRL was originated in 1969, the primary thrust of our measurement and evaluation research program has been directed toward the measurement aspect of the criterion problem. Resources committed to this task have been quite limited, due primarily to other commitments within the Division such as development of the Advanced Instructional System. Rarely has more than one man-year been devoted to measurement/evaluation. Within these constraints, we have tried to be responsive to the immediate needs of the Air Force as well as to investigate new techniques for incorporation into computer based instructional systems.

During the 1969-1972 time period, problems of measuring student attitude and student achievement occupied our attention. The attitude measurement project attempted to develop a new Student Critique Form for potential ATC usage. A series of reports (7, 8, 9) was issued covering the development of the critique scales, the formation of norm groups, scale reliability, factor analysis of the questionnaire, and use of the discriminant function to support item validity. The norm referenced approach described by our researchers in the final report (12) was judged by ATC personnel to be operationally infeasible; consequently, the newly developed critique form was never used.

In the achievement domain, we investigated the utility of confidence testing in an Air Force environment (2, 3, 4, 5). Confidence testing is a technique for test scoring, where students are asked to express the degree of confidence they have in their answer. Confidence testing could increase the predictive validity of test scores in two ways: (a) by making constructive use of partial knowledge in determining an examinee's true score, and (b) by reducing test anxiety. Of the available techniques for allocating confidence, two methods were studied in the classroom (6). Neither proved superior, and the students were relatively indifferent to use of either technique. The most serious objection came from instructors who felt that the system was too complex to score by hand. However, the results of this study may one day be applied through incorporation into a computer scoring routine.

By 1972, we had turned our attention to finding alternatives to the multiple choice format for testing the knowledge domain and to the development of more sensitive scoring systems (1, 14). This effort culminated in a study by Siegel et al. (15) in which several advanced measurement techniques were tried in a classroom setting. Included were novel item formats such as analogies, pictorial testing, and cognition of figural systems as well as new scoring methods such as confidence testing, sequential testing, and theory of signal detection. Though these techniques were, on the whole, successfully demonstrated in the study, they were not adapted on a wide scale, probably because ATC first-line evaluation personnel were not trained in their use.

In search of more efficient ways of measuring an examinee's knowledge and skills, we initiated work in adaptive or "tailored" testing. Here, a reduced set of items is given to an examinee, dependent on his or her previous pattern of responses. Our initial efforts in adaptive testing were to consider the issues involved in implementing this technique in a computer based training system (10). Waters (17) also conducted an empirical investigation of one approach--the Stradaptive model for measuring ability--and concluded that the model held promise.

Hansen et al. (11) successfully implemented two adaptive testing algorithms, Flexilevel and Hierarchical, in the Precision Measuring

73

Equipment Specialist course at Lowry. Results from this study are decidedly encouraging. Time savings approximated 20%, and accuracy of measurement was nearly identical to conventional procedures.

A 1974 study (16), which surveyed ATC measurement/evaluation procedures in the context of the ISD model, developed some information which laid the groundwork for our current interest in the definition aspect of the criterion problem. An in-house follow-on study (13) appraised the ATC graduate evaluation system, presented a method for determining over- and under-training, and suggested a task clustering approach to linking job performance with training objectives.

Present Work

Work on adaptive testing has been undertaken primarily to decrease test time. In a well described instructional sequence, frequent measurement yields assurance that the student has attained prerequisite basic concepts and skills before proceeding to more complex areas in the curriculum. However, no single model or algorithm for adaptive testing has a clear lead at this time, nor are any ready for widespread implementation. More work needs to be done particularly in the theoretical development of adaptive criterion referenced performance tests. Consequently, we are participating in an interservice project which is supporting work in this area by Dr. David Weiss at the University of Minnesota. Another basic research contract with the same general objective, although with a somewhat different approach, is also being supported.

Development of an Advanced Field Evaluation System for ATC represents our first real attempt to validate the link between training performance and job performance and addresses the criterion definition aspect. While the primary purpose of the research is to provide more useful information about training adequacy, a by-product of this study will be a direct check, independent of the occupational survey reports, on whether tasks trained are actually performed on the job. Hopefully, as well, there will emerge a more sensitive scale or measure of job performance. Another procedure that we have investigated for increasing testing efficiency is called symbolic performance testing. The underlying concept in this technique is to capture the essential features of a performance test in either a paper-and-pencil mode or by means of audiovisual or computer graphic presentation.

Thus, we administer a symbolic version, or analog, which correlates very highly with the actual performance test. In the process, we avoid consuming instructor and equipment time for test purposes and can deal with more than one or two students at a time. We are currently working on a demonstration of this technique in an electronics training course at Lowry AFB. Previous work on symbolic performance testing has not been very encouraging. Nevertheless, the potential increase in testing efficiency makes continued exploration of symbolic performance testing worthwhile.

Since the advent of criterion referenced measurement, one of the major tools used by the ATC instructor has been the criterion checklist. Because accurate measurement requires reliability as a precondition, we are currently investigating the reliability of this device in two ATC courses. We hope to be able to suggest operational practices which would increase the reliability of measurement from use of criterion checklists.

Future Research

Requirements validation, referred to in Table 1, is meant to encompass research to ensure that training objectives flow from job requirements. We would agree that some theory, concepts, skills, or abilities should be taught, even though these do not appear to be job requirements *per se*. The object here would be to discover better ways of judging which enabling objectives are prerequisites to job performance and which are irrelevant. The student himself may be a fruitful, but often overlooked source of ideas, and so we are led full circle back to student critiques as a method for developing this information.

Returning to the measurement aspect, we intend to pursue applications of latent trait theory to ATC measurement problems. Latent trait theory is a relatively new approach to measurement. Popularized by Lord (1952, 1953a, 1953b), latent trait theory has the potential to help solve many criterion-related measurement problems. Hambleton et al. (1977) cite the disadvantages of classical measurement procedures; among these are sample scientific item parameter estimates, and the fact that they have no utility in determining how a given examinee will perform on a particular item or set of items. Latent trait theory permits us to predict item performance by individual examinees based upon the underlying trait or characteristic being measured. It does not rely on the classical "standard error of measurement" which assumes that examinees of all ability levels have equal errors of measurement. Latent trait theory also allows us to get away from the concept of using only group correlation coefficients between predictor and criterion test scores to determine the utility of a measurement procedure. The concept of "test information function" allows us to compare different instruments or procedures in terms of the relative amount of information which they produce about the underlying trait.

A relatively large and growing amount of research has been done in the use of latent trait parameter estimates for selection and classification. Aside from the on-going work by Weiss and his associates at the University of Minnesota, practically no research has been done in an instructional environment. We plan to look at such applications in the near future, probably in our FY79 program.

One of our major concerns is the effect of having a multi-dimensional instructional situation as opposed to a relatively uni-dimensional aptitude measurement problem. As current latent trait models are defined, a uni-dimensional latent trait is assumed. We must either examine the robustness of existing models to violation of this assumption or create new, more complex, models which can handle multi-dimensional data. If one of these alternatives proves fruitful, many of the scaling, sampling, and lack of individual predictive problems with conventional criterion predictions may be eased.

Continued work on adaptive testing is a clear future direction. We would propose to implement those models which survive our present studies and meet the tests of practicality, ease of use, efficiency, and accuracy.

Still a third aspect to the criterion problem, not considered so far, is the question of how to set cutoff scores on test instruments. What rationale should be used for deciding mastery level? A further factor is the utility or cost of testing. Is the information gained from the test worth the cost of administration? If cost were taken into account, perhaps we would conclude that the test ought not to be given at all! A decision theory approach (Cronbach & Gleser, 1965) based on the notion of utility may be fruitful for investigating these two additional aspects.

Figure 2 is a graphic representation of the problem one faces in setting cutting scores on a test. In a roughly normal distribution of test scores, students tend to fall into three discernible groups: masters, non-masters, and a fairly large middle group which has test scores between C_1 and C_2. One would need to collect more information about this middle group to render an effective mastery decision. This may be uneconomical in certain instances. If C_2 is chosen as the cutting score, the shaded area represents errors of classification on the task. With C_2 as the cutting score, false positives are quite small and false negatives relatively large; the opposite is true if C_1 is chosen.

A decision-theoretic way of thinking may be helpful in setting the cutting score. The importance of the decision being made dictates whether C_1 or C_2 is the most beneficial place for the cutting score. If the consequences of task success and failure can be quantified in a dollar metric, the costs of additional testing could be combined for various cutting levels, and a more rational decision could be made. These questions remain to be explored in future research.

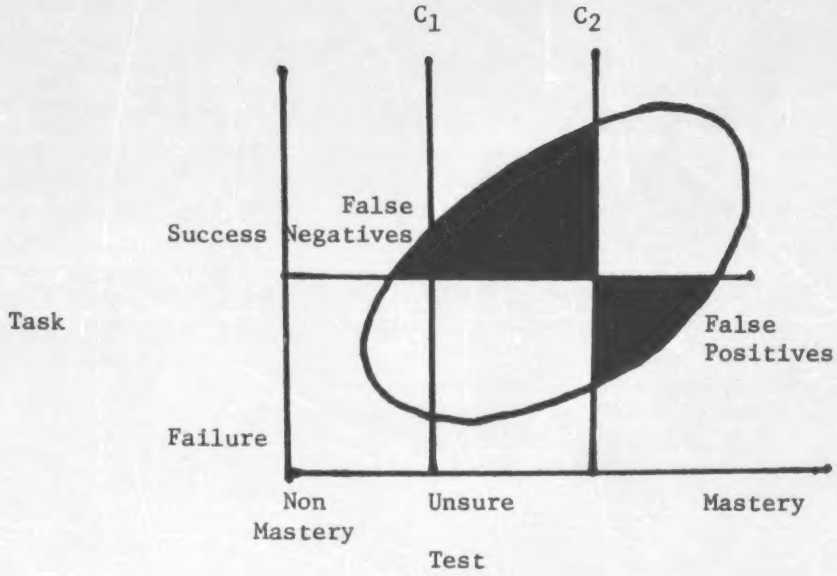


Figure 2. Cutoff scores and decision errors.

Conclusion

In summary, it should be emphasized that for the training community progress on the criterion problem will come when both the measurement and definition aspects have been addressed. We have shown to what extent our program has been concerned with these issues. More work needs to be done on the definition aspect in order to assure ourselves that job relevant behaviors are being trained in an effective manner.

Much of what has been presented in this paper is clearly applied, even "action-oriented," research. That is, known techniques are applied to solve operational problems. We have a strong bias in this direction and feel that such is a proper orientation for a military R&D organization. However, some emphasis on theoretical development, advances in statistical methodology, and innovation in measurement techniques will be maintained. We must continue to support and encourage basic research so that new tools will be available to solve problems yet unstated.

We hope to have learned some lessons in our 8-year existence. Many of these are not research lessons but guidelines for translating our research into operational programs. We must constantly be alert for closer coordination with our users, not only to be responsive to real needs, but also to help with the problem of personnel turnover and changing perceptions of needs. In addition, we must provide transition plans to include support and training where needed so that improvements may be institutionalized, for instutionalization of our research must be the overriding goal of our measurement/evaluation program.

References

Cronbach, L.J., & Glaser, G.C. Psychological tests and personnel decisions. Urbana IL: University of Illinois Press, 1965.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., & Gifford, J. Developments in latent trait theory: A review of models, technical issues, and applications. Paper presented at a joint meeting of NCME and AERA, New York, 1977.

Lord, F.M. A theory of test scores. Psychometric Monograph, No. 7, 1952.

Lord, F.M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953a, 18, 57-75.

Lord, F.M. The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 1953b, 13, 517-548.

Thorndike, R.L. Personnel selection. New York: Wiley, 1949.

BIBLIOGRAPHY

(1) Bergman, B.A., & Siegel, A.I. Training evaluation and student achievement measurement; a review of the literature (AFHRL-TR-72-3, AD-747 040). Lowry AFB CO: Technical Training Division, January 1972.

(2) Boldt, R.F. A simple confidence testing format (AFHRL-TR-71-31, AD 737 113). Lowry AFB CO: Technical Training Division, July 1971.

(3) Boldt, R.F. An approximately reproducing scoring scheme that aligns random response and omission (AFHRL-TR-74-99, AD-A005 301). Lowry AFB CO: Technical Training Division, November 1974.

(4) Echternacht, G.J. Use of confidence testing in objective tests (AFHRL-TR-71-32, AD-734 031). Lowry AFB CO: Technical Training Division, July 1971.

(5) Echternacht, G.J., Boldt, R.F., & Sellman, W.S. User's handbook for confidence testing as a diagnostic aid in technical training (AFHRL-TR-71-34, AD-731-192). Lowry AFB CO: Technical Training Division, July 1971.

(6) Echternacht, G.J., Sellman, W.S., Boldt, R.F., & Young, J.D. An evaluation of the feasibility of confidence testing as a diagnostic aid in technical training (AFHRL-TR-71-33, AD-734 032). Lowry AFB CO: Technical Training Division, July 1971.

(7) Federico, P.A. Development of psychometric measures of student attitudes toward technical training: reliability and factorial validity (AFHRL-TR-70-37, AD-723 314). Lowry AFB CO: Technical Training Division, November 1970.

(8) Federico, P.A. Identifying item validity indices utilizing a multi-variate model (AFHRL-TR-71-16, AD-729 763). Lowry AFB CO: Technical Training Division, April 1971.

(9) Federico, P.A. Degree of evaluative assertions ascribed to an attitude universe as a function of measurement format (AFHRL-TR-71-8, AD-736 788). Lowry AFB CO: Technical Training Division, December 1971.

(10) Hansen, D.N., Johnson, B.F., Fagan, R.L., Tam, P., & Dick, W. Computer-based adaptive testing models for the Air Force technical training environment phase I: Development of a computerized measurement system for Air Force technical training (AFHRL-TR-74-48, AD-785 142). Lowry AFB CO: Technical Training Division, July 1974.

(11) Hansen, D.N., Harris, D.A., & Ross, S. Flexilevel adaptive testing paradigm: Validation in technical training (AFHRL-TR-77-35 (1)). Lowry AFB CO: Technical Training Division, Air Force Human Resources Laboratory, July 1977.

(12) Miller, G.G., & Sellman, W.S. Development of psychometric measures of student attitudes toward technical training, norm group report (AFHRL-TR-73-15, AD-775 151). Lowry AFB CO: Technical Training Division, October 1973.

(13) Pennell, R., Harris, D., & Schwille, J. Appraisal of Air Force Training Course Field Evaluation System (AFHRL-TR-76-63). Lowry AFB CO: Technical Training Division, October 1976.

(14) Siegel, A.I., Bergman, B.A., Federman, P., & Sellman, W.S. Some techniques for the evaluation of technical training courses and students (AFHRL-TR-72-15, AD-753 094). Lowry AFB CO: Technical Training Division, February 1972.

(15) Siegel, A.I., Bergman, B.A., & Miller, G.G. Adaptation of advanced measurement and evaluation techniques for utilization in Air Force technical training systems (AFHRL-TR-73-18, AD-773 802). Lowry AFB CO: Technical Training Division, November 1973.

(16) Siegel, A.I., Federman, P.J., & Sellman, W.S. A survey of student measurement and course evaluation procedures within the Air Training Command (AFHRL-TR-74-5, AD-786 041). Lowry AFB CO: Technical Training Division, July 1974.

(17) Waters, B.K. Empirical investigation of the stradaptive testing model for the measurement of human ability (AFHRL-TR-75-27). Williams AFB AZ: Flying Training Division, October 1975.

X

OVERVIEW OF ADVANCED SYSTEMS DIVISION CRITERION RESEARCH (MAINTENANCE)

John P. Foley, Jr.
Advanced Systems Division
Wright-Patterson Air Force Base, Ohio

Introduction

The Advanced Systems Division (AS) of Air Force Human Resources Laboratory (AFHRL) has had two separate and distinct criterion R&D programs--one concerning pilot performance, and the other concerning maintenance performance. Today I am addressing our maintenance program.

Maintenance of hardware is currently an extremely costly operation for the Department of Defense (DoD). High maintenance cost is the primary cause of high systems ownership cost. For some electronic maintenance specialties, nearly 1 year of broad formal training is given first enlistment personnel. And maintenance training generally is long and costly. Even with such lengthly training, the efficiency of maintenance could be greatly improved. Improved job instructions and information, as well as increased use of job (task) oriented training have great potential for decreasing maintenance training time[1] and improving the job performance of maintenance tasks.

But, to maximize such potential and to ensure more efficient maintenance, the criteria for the selection, training, assignment, and promotion of maintenance men should be the demonstrated ability of maintenance personnel to perform the tasks of their jobs. To enforce such criteria, the key job tasks must be identified and the ability to perform identified tasks must be ascertained. Since the ability to perform many or most of the identified tasks will not be part of the normal repertoire of those being selected for jobs, appropriate action must be taken to develop the ability to perform job tasks. Of course, these actions are "easier said than done."

The Criterion Problem

If we can produce a measuring device that actually measures the ability to perform the desired behaviors under all the desired conditions, we have an ultimate criterion measure. But the fact that we usually cannot develop such a device forces us to settle for a secondary criterion measure which is, at best, somewhat different than the ultimate. As we see it, this difference between the real world and the simulation of the real world for testing purposes is the criterion problem.

A common example of such a criterion problem presents itself when we attempt to measure an individual's ability to drive automobiles. To measure such ability completely, we would have to devise a test that would measure his ability to perform all driving tasks of all automobiles, on all types of roads, in all traffic conditions, under all types of weather conditions, whether he is being observed or not. It is obvious that it would be virtually impossible to meet all of these conditions under practical testing conditions. We, therefore, settle for a less rigorous test criterion. We assume that he can drive any automobile adequately, if he demonstrates in a performance test that he can perform most driving tasks in one automobile, in normal traffic, while being observed.

But many times, it is inconvenient and considered too costly to administer even such a driver performance test, and an attempt is made to develop a paper-and-pencil test which will determine that an individual can drive adequately. But such a test cannot be considered to be a valid substitute unless a high empirical relationship to the criterion measure can be demonstrated. In the practical world of test development, the driver performance test would be considered an adequate, near ultimate criterion test for validation of such a paper-and-pencil substitute. Many times such a paper-and-pencil test is used without being validated against such a near ultimate criterion test. The use of such an unvalidated test would be an extremely dangerous practice, since it is assumed by most users that it measures an individual's ability to drive, when in fact, we are not sure what it is measuring.

This criterion problem has long plagued measurement theorists and practitioners, as well as curriculum researchers. The use of job tasks, and performance examinations based on these tasks as near ultimate criteria for evaluation of selection devices, was first emphasized as a result of the work of Army and Navy measurement psychologists during World War II. In 1946, Jenkins discussed the problem in light of the experiences of Navy psychologists in an article in the American Psychologist, entitled "Validity for what?"

> Psychologists in general tended to accept the tacit assumption that criteria were either given of God or just to be found lying about. . . . The novice of 1940, searching through many textbooks and much journal literature, would have been led to conclude that expediency dictated the choice of criteria and that the convenient availability of a criterion was more important than its adequacy.

In 1964, the late Rains Wallace presented a paper at the annual convention of American Psychological Association (APA), which also appeared in the American Psychologist (Wallace, 1965a). It indicated that much of what Jenkins said in 1946 was still true.

> In the 18 years which have followed, we have become wiser and sadder about the criterion problem. If we have not accomplished a great deal, if we tend to use the expedient criterion with the comforting thought that some day we will get down to constructing better ones, if we concentrate on criteria that are predictable rather than appropriate, we do operate with varying levels of guilt feelings. We have not done much about it, but we know we should.

In 1965, Wallace presented another paper in which he addressed the criterion problem very succinctly as it applies to electronic maintenance.

> All of this is prelude to my main thesis which is in no sense revolutionary, original, or controversial. I state it because it is honored in the breach. It is that the nature of our proficiency measures determines how we select, classify, train, maintain, and assess our human resources. If the measures are largely irrelevant to the jobs we want done, we will select the wrong men, classify them incorrectly, and train them wrong. This is true because these proficiency measures are, or should be, the criteria against which we validate our selection and classification procedures and evaluate our training content and methodology or our supervisory techniques. Thus, if I use a test of advanced electronics theory as the proficiency measure for electronics maintenance and as the criterion against which to evaluate a test for selecting men to go into maintenance training, I will end up choosing a selection test which rejects men who are not well above average in both reading and arithmetic ability. In the process, I might reject a great many who are outstanding in their ability to get their hands on a piece of machinery and make it work. I might also accept a number who (like myself) are so lacking in the simplest manipulative ability that their hands could have been cut off at the wrists at birth without seriously affecting their outputs. So, when I decided what proficiency measures to use, I also decided what kind of men I was going to put into training for the job.
>
> But it doesn't end there. For when I now approach the problem of how to train men to perform the tasks involved in the job, I must make decisions about what should be taught and what methods should be used in teaching it. The only way I have of reaching such decisions (except by divination which is, admittedly, not a rare procedure) is to measure and compare

> the performance achieved with various curricula and methodologies. So, in the case of the electronics maintenance course, I put in lots of reading about electronics theory and I produce graduates who can read and write electronics theory while their equipment deteriorates in hopeless inoperativeness (Wallace, 1965b, p. 4).[2]

Influenced in part by the above statement, we at the Advanced Systems Division decided to do something about the criterion problem as it applied to maintenance. And, although our work was at times delayed and sidetracked, 12 years later we do have some R&D completed which we can talk about. However, the grim and vivid picture that Rains Wallace painted in 1965 is still true for most of the operational Air Force.

Our approach to the criterion problem has been to study and analyze both measurement literature and maintenance jobs, and to develop job task performance tests (JTPT) for key maintenance tasks which were selected on the basis of these analyses. We developed these JTPT to be as near to ultimate job criteria as possible in keeping with the following suggestion of Frederiksen:

> The objective, presumably, is to get as close as is feasible to the ultimate criterion; but as has just been seen, when one gets too close to the real-life situation, control of the conditions for adequate observation is lost. Observation of real-life behavior is ordinarily not a suitable technique for measurement. The type of measure that is recommended for first consideration in a training evaluation study is the type which most closely approximates the real-life situation, that which, in this chapter, has been called eliciting lifelike behavior. If it is not feasible to wait for the behavior to happen in real life, then lifelike occasions can be provided for the behavior to occur in a test situation (Frederiksen, 1962, p. 334).

Admittedly, an examination made up of tasks removed from their actual job environment is not an ultimate criterion test. Under actual job situations, the graduate may have to perform these tasks in cramped quarters; under stresses of time, noise, heat, or cold; or with an excited boss interfering. These conditions of stress are usually not constant variables, but change from day-to-day and from hour-to-hour. The assumption usually has to be made that the individual can perform a task under conditions of stress, provided he can perform the same task well under normal conditions. A formal performance examination has its own set of stresses, which may not be the same as job stresses, but their presence may tend to offset the lack of job stresses. Formal job task performance examinations are the closest usable simulation of the real maintenance jobs presently available. They are far better than no performance tests at all.

Review of Performance Measurement (PM) Literature

In regard to the literature reviews and analyses made for PM (Foley, 1967, 1974), many valuable PM efforts have been reported by the Army, Navy, and Air Force. However, most of these efforts have not been systematic efforts, having as their prime objective the improvement of the state-of-the-art of PM. Rather, they have been ad hoc PM developments to support job oriented training research programs. A notable exception was the work of the Air Force Personnel and Training Center (AFPTRC) Maintenance Laboratory. (Another more recent systematic Army effort, accomplished by the Human Resources Research Organization (HumRRO) was not covered in these reviews (Vineberg, Taylor, & Caylor, 1970a, 1970b; Vineberg & Taylor, 1972a, 1972b)). As to civilian R&D, during the initial PM literature review (Foley, 1967), a serious attempt was made to identify and include the results of PM R&D from the civilian vocational education establishment. None was found.

A substantial outcome of the review of other PM efforts was a consolidation of research results concerning the correlations between results of PM for various maintenance tasks and paper-and-pencil theory tests, job knowledge tests, and school marks. As to their value for measuring ability to perform maintenance tasks, this research evidence gives a low rating to all of these paper-and-pencil based measures of school and job success. Table 1 shows correlations that have been obtained by comparing JTPT to theory tests, and to job knowledge tests. The latter two are paper-and-pencil tests. Table 1 also includes correlations of JTPT with school marks. As indicated earlier school marks have been heavily weighted with the paper-and-pencil test scores. An examination of this table indicates that the correlations of JTPT scores with theory test scores are generally somewhat lower than with job knowledge tests. None of these measures is sufficiently valid for use as substitutes for JTPT (Foley, 1967, 1974).

The personnel system, which includes formal training, depends almost exclusively on such paper-and-pencil tests for making initial selection, for ascertaining effectiveness of training, and for the promotion of maintenance personnel. The effectiveness of formal training for the mechanical maintenance specialties is measured mainly by scores obtained from such paper-and-pencil job knowledge tests, even though the students in these training programs have received at least some "hands-on" practice on many mechanical maintenance tasks. The measures of effectiveness of formal training programs for the electronic maintenance specialties include scores from paper-and-pencil job knowledge tests, as well as theory tests. Students in these electronic maintenance courses receive little if any "hands-on" practice in their maintenance tasks.

Table 1. Correlations Between Job-Task Performance Tests and Theory Tests, Job Knowledge Tests, and School Marks

Researchers	Type of Job Task Performance Tests (JTPT)	Theory Tests	Job Knowledge Tests	School Marks
Anderson (1962a)	Test Equipment JTPT			.18-.33
Evans and Smith (1953)	Troubleshooting JTPT	.24 & .36	.12 & .10	.35
Mackie et al. (1953)	Troubleshooting JTPT	.38		.39
Saupe (1955)	Troubleshooting JTPT		.55	.55
Brown et al. (1959)	Troubleshooting JTPT	.40		
	Test Equipment JTPT		.29	
	Alignment JTPT		.28	
	Repair Skills JTPT		.19	
Williams and Whitmore (1959)	Troubleshooting JTPT			
	(Inexperienced subjects)	.23		
	(Experienced subjects)	.15		
	Adjustment JTPT			
	(Inexperienced subjects)	.02		
	(Experienced subjects)	.21		
	Acquisition Radar JTPT			
	(Inexperienced subjects)	.03	.36	
	(Experienced subjects)	.14	.22	
	Target Tracking Radar JTPT			
	(Inexperienced subjects)	.24	.33	
	(Experienced subjects)	.20	.38	
	Missile Tracking Radar JTPT			
	(Inexperienced subjects)	.09	.15	
	(Experienced subjects)	.19	.32	
	Computer JTPT			
	(Inexperienced subjects)	.08	.24	
	(Experienced subjects)	.06	.14	
	Total JTPT			
	(Inexperienced subjects)	.14		
	(Experienced subjects)	.20		
Crowder et al. (1954)	Troubleshooting JTPT	.11	.18-.32	

The selection tests for both mechanical and electronic maintenance specialties have been standardized against composite scores from paper-and-pencil tests. This means that the people selected for the maintenance specialties have been selected not on their aptitude for performing the tasks of their maintenance jobs, but on their aptitude for making high scores on paper-and-pencil, theory, and job knowledge tests.

Our specialty knowledge test (SKT) and the promotion fitness examination (PFE) used for advancement up the maintenance career ladders also are paper-and-pencil job knowledge tests. At the present time, throughout his whole career, a maintenance specialist is not required to demonstrate on formal JTPT that he can efficiently and effectively perform the tasks of his job.

The Man-Machine Interface for Maintenance

The maintenance R&D supported by AS has emphasized the man-machine interface. From this point of view, PM for all personnel associated with machine systems must determine the ability of such personnel to perform tasks generated by the man-machine interface. Although there may be some overlap, most of the task functions demanded by a machine system of its operator personnel are different from those task functions demanded of its maintenance personnel. Herein lies most of the unique, distinguishing characteristics of PM for maintenance. As a result, this section of my paper will be devoted to a discussion of the complexity of maintenance task functions.

Past Human Factors Emphasis

But before discussing the characteristics of task functions for maintenance, it might be well to call attention to the fact that human factors establishments have given much more attention to the operator interface with machines than to the maintenance personnel interface. Many actions are taken to maximize effective and efficient performance of the operator. Work stations are human engineered to maximize the efficiency and comfort of the human operator. Major training facilities are provided so that operators can receive a large amount of supervised practice in performing typical tasks of their job. Graduation from training is based primarily on demonstrated ability to perform job tasks. And, periodic checks are made of the operator's ability to perform the critical tasks of his job. These, of course, are not all of the many efforts made to maximize the performance of human operators.

Generally, the human factors establishment has given little attention to the effectiveness and efficiency of the maintenance man's interface with hardware. The maintenance work of AS, including the PM work, has emphasized this neglected interface, but typically, this part of our program has received little management visibility or support.

The Structure of the Man-Machine Interface for Maintenance

One of the results of our R&D for maintenance has been the evolution and articulation of a structure for handling maintenance functions and their complex relationships in a systematic manner. This structure includes (1) standard maintenance functions and action verbs, (2) a working definition of a maintenance task, and (3) schemes for handling the complexities of maintenance tasks.

Standard Maintenance Functions and Action Verbs

The establishment of standard maintenance functions and actions verbs has been one of the widely accepted results of the Air Force Systems Command's (AFSC) job performance aids (JPA) effort entitled "Presentation of Information for Maintenance and Operation" (PIMO). (Although the PIMO project was managed by the Space and Missile Systems Organization (SAMSO) of AFSC, AS provided active participation and technical inputs during the entire project from 1966 through 1969. AS has incorporated the key findings and outputs of PIMO in its own JPA efforts.) Early in the PIMO project, it was found that many maintenance action verbs and functions were used by maintenance people, some with several different meanings. Part of this confusion was caused by the language used in maintenance technical orders which were written by different people and produced by many different hardware manufacturers. As a result, maintenance technicians themselves did not generally use precise language. A study was made to identify and define these action verbs. Where two or more verbs were used to indicate a similar action, the preferred verb was selected based on the expressed preferences of a sample of maintenance men with a wide range of maintenance Air Force Specialty Codes (AFSCs). The use of the preferred verbs of this list is now a firm requirement of Air Force technical order specifications, as well as of recent Army and Navy specifications (see Joyce, Chenzoff, Mulligan, & Mallory, 1973, pp. 97-142).

A Working Definition of a Maintenance Task

Within this list of action verbs are a number of key action verbs (functions). A key action verb, with an appropriate specific hardware unit as its predicate, becomes a task statement. Such a task statement represents a maintenance task which can be demanded by the existence and operation of a specific machine subsystem. A list of these functions is found in AFHRL-TR-73-43(I) (Joyce et al., 1973, pp. 19, 20). This list includes functions which are found in both mechanical and electronic jobs. Some apply to only mechanical jobs and some apply to both.

Schemes for the Systematic Consideration of Maintenance Functions and Tasks

Three schemes have been developed for the systematic consideration of maintenance functions and tasks, and the key factors that affect them.

Scheme One--A convenient model for categorizing these maintenance functions with relation to the type of hardware and the level of maintenance is presented in Figure 1. The common maintenance functions already mentioned together with the usage of test equipment and hand tools are represented on one axis of the model. Since mechanical and electronic subsystems usually require a different variety of maintenance actions, they are represented by another axis. (In regard to this axis, mechanical maintenance could be further divided into two categories, one represented by hardware such as jet engines, and another by hardware such as airframes and tank and ship hulls.)

The third axis of the model represents the three levels or categories of maintenance now found in the military services. Organizational maintenance is the first level. It is usually aimed at checking out a whole machine subsystem and correcting any identified faults as quickly as possible. Flight line maintenance falls in this category. A system is checked out. If it does not work, the line replaceable unit (LRU) or "black box" causing malfunction is identified and replaced. This major component is then taken to the field shop (intermediate maintenance) where it is again checked out and the faults, authorized for correction, are corrected. The corrective actions, authorized at the intermediate level, vary greatly from system to system depending on the maintenance concept of each system. On some systems, the maintenance man will troubleshoot the "black box" to the piece part level. In more modern equipment, he will identify a replaceable module made up of many piece parts. Some modules are thrown away, others sent to the depot for repair. Any line replaceable units which the field shop are unable, or unauthorized, to repair are sent to the depot for overhaul.

Organizational and intermediate level organizations are manned primarily by enlisted technicians whose average length of service is rather short (slightly more than 4 years in the Air Force). Depots are manned largely by civilian personnel with a much higher level of experience and longer retention time. Using this model, it has been possible to specify areas of concentration for study.

Since PM requirements for maintenance are so different for the various blocks indicated in this model, it is extremely important that PM researchers indicate the precise blocks of their concentration. To date, AS has concentrated on the shaded electronic portions of this model (Figure 1). The resultant model battery of 48 JTPT together with their symbolic substitutes will be described later. In addition, a battery of 11 JTPT was developed on an ad hoc basis (Shriver & Foley, 1975) for mechanical tasks at the organizational level of maintenance (see shaded portion of Figure 2). The HumRRO work, mentioned previously (Vineberg et al., 1970a, 1970b; Vineberg & Taylor, 1972a, 1972b) was concerned with mechanical hardware (tank and truck). The 13 tests developed concerned the maintenance functions which are indicated by the shaded portions of Figure 3.

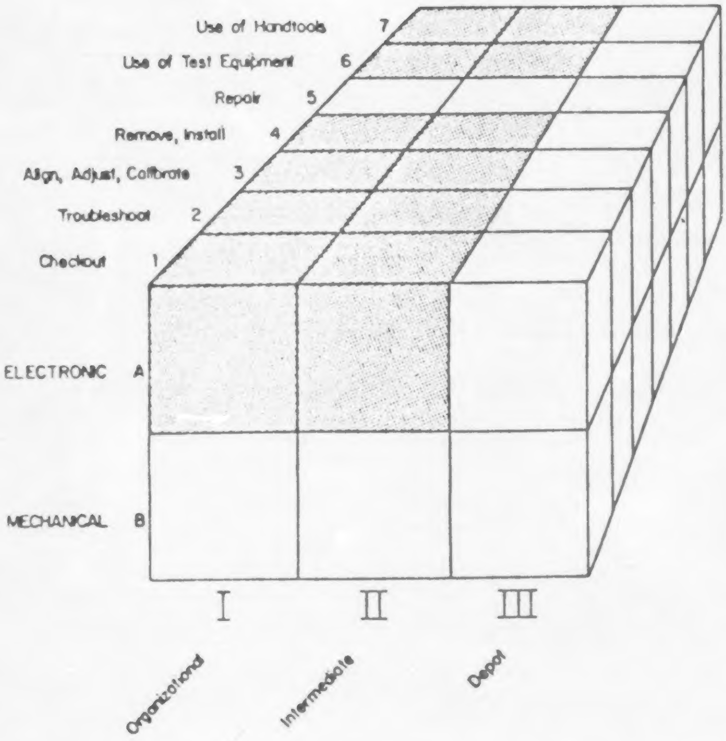

Use of Handtools 7
Use of Test Equipment 6
Repair 5
Remove, Install 4
Align, Adjust, Calibrate 3
Troubleshoot 2
Checkout 1
ELECTRONIC A
MECHANICAL B
I
II
III
Organizational
Intermediate
Depot

Figure 1 - A functional representation of the DOD Maintenance Structure (Shaded portion indicates scope of AFHRL PM development for electronic maintenance).

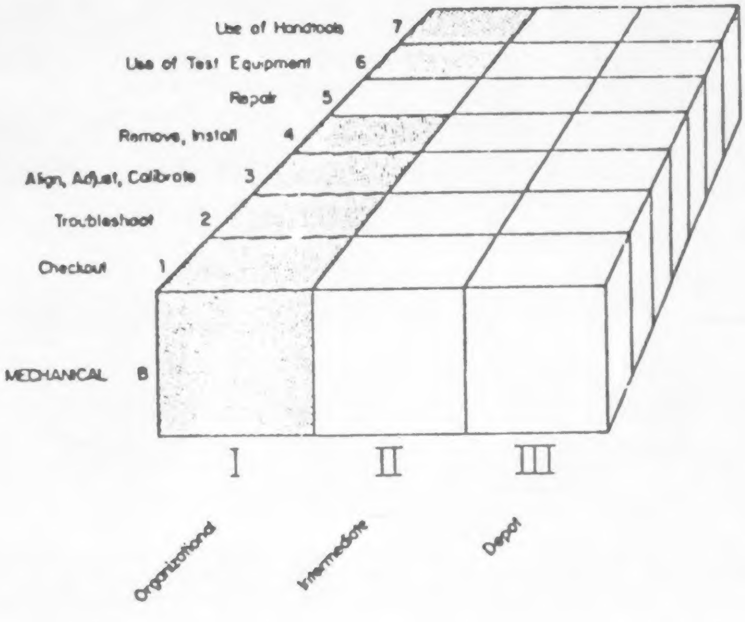

Use of Handtools 7
Use of Test Equipment 6
Repair 5
Remove, Install 4
Align, Adjust, Calibrate 3
Troubleshoot 2
Checkout 1
MECHANICAL 8
I
II
III
Organizational
Intermediate
Depot

Figure 2 - A functional representation of the scope of AFHRL PM development for mechanical maintenance.

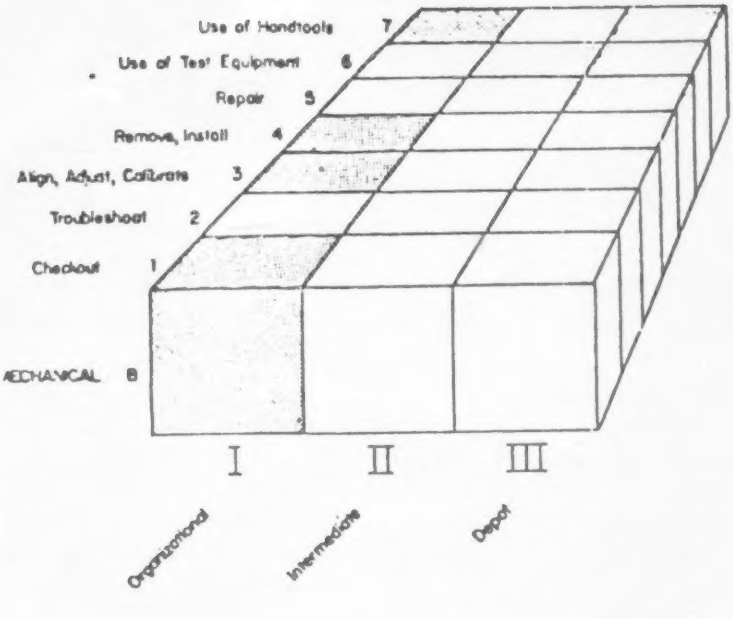

Use of Handtools 7
Use of Test Equipment 6
Repair 5
Remove, Install 4
Align, Adjust, Calibrate 3
Troubleshoot 2
Checkout 1
MECHANICAL 8
I
II
III
Organizational
Intermediate
Depot

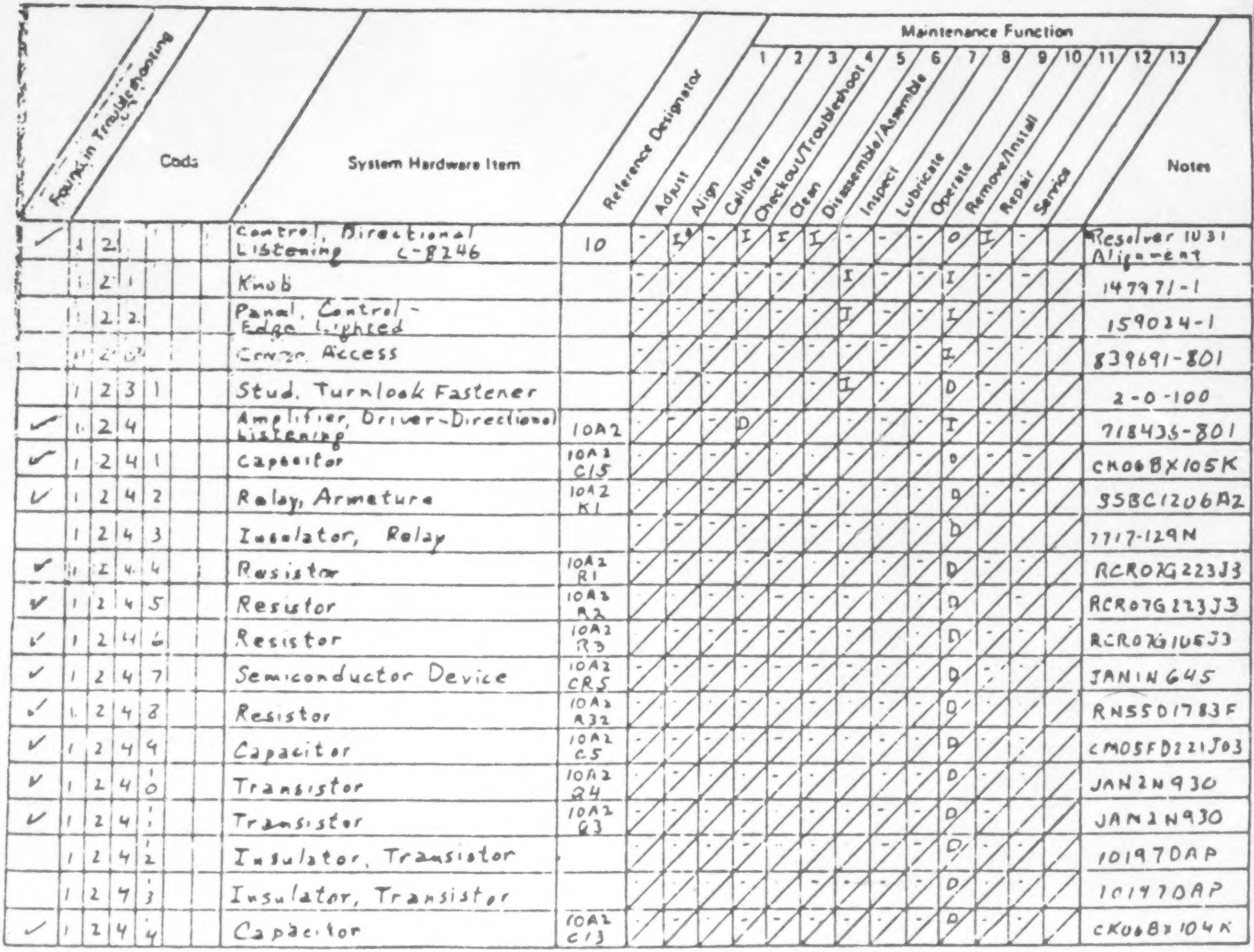
Figure 3 - A functional representation of the scope of the HumRRO PM development for mechanical maintenance (Vineberg et al, 1970b).

Scheme Two--Maintenance functions have limited meaning unless applied to specific hardware. A task identification matrix (TIM) is an extremely effective and necessary device for interfacing these maintenance functions with the appropriate hardware units and thus identifying the maintenance tasks that are generated by a specific machine subsystem (see Figure 4). The TIM, when properly structured, will reflect the maintenance level or levels of interest, that is organizational, intermediate and/or depot. AFHRL-TR-73-41(I) (Joyce et al., 1973, pp. 16-37) provides detailed directions for developing a TIM.

Scheme Three--A matter of serious concern when developing and structuring PM for maintenance tasks is the interaction among the maintenance tasks for one hardware. A four-level hierarchy of dependencies can be stated. Figure 5 gives a graphic presentation of these dependencies among maintenance activities for an electronic hardware.

The checkout of the AN/APN-147 (Doppler Radar), for example, can be a task in its own right. But the same checkout activity becomes an element of other major tasks such as calibrate. The calibration of doppler radar includes the operation of specific general and special test equipments, the use of specific hand tools, as well as the checkout activity. Troubleshooting of an electronic equipment, such as AN/APN-147, requires the use of general and special test equipments. It may require remove and install activities and/or adjust, align, and calibrate activities. Efficient troubleshooting practice usually requires the use of a cognitive strategy to adequately track the dependent activities (but the cognitive strategy in itself is not troubleshooting). Any troubleshooting task should begin and end with an equipment checkout. Because of these various and varying dependency relationships, such activities as checkout, remove, install, disassemble, adjust, align, calibrate, or troubleshoot cannot legitimately be considered as discrete tasks, even for one electronic system.

Another confounding factor is the false correspondence that the same functional verbs create when applied to different electronic hardware. For example, personnel with the Avionic Inertial and Radar Navigation Systems Specialist, AFSC 328X3, are maintaining at least 50 major electronic subsystems. Many vintages of hardware design are represented. The checkout activity for each is different (both in content and difficulty) and in some cases, very different. The lack of correspondence of alignment, calibration, and troubleshooting tasks from one specific equipment to another is even greater. An example of the lack of correspondence from one hardware to another is the wide difference in the content and difficulty of troubleshooting tasks between two doppler radars. The AN/APN-147, which is used on the C-130 and C-141, has approximately 14,000 shop replaceable units (SRU) whereas the inertial doppler navigation equipment (IDNE) on the C-5 has only 28. This lack of correspondence of functions across electronic hardware



Found in Troubleshooting	Code						System Hardware Item	Reference Designator	Maintenance Function													Notes
									1 Adjust	2 Align	3 Calibrate	4 Checkout/Troubleshoot	5 Clean	6 Disassemble/Assemble	7 Inspect	8 Lubricate	9 Operate	10 Remove/Install	11 Repair	12 Service	13	
✓	1	2					Control, Directional Listening C-8246	10	-	I	-	I	I	I	-	-	-	O	I	-		Resolver 1U31 Alignment
	1	2	1				Knob		-	-	-	-	-	-	I	-	-	I	-	-		147971-1
	1	2	2				Panel, Control - Edge Lighted		-	-	-	-	-	-	I	-	-	I	-	-		159024-1
	1	2	3				Cover Access		-	-	-	-	-	-	-	-	-	I	-	-		839691-801
	1	2	3	1			Stud, Turnlook Fastener				-	-		-	I	-	-	D	-	-		2-0-100
✓	1	2	4				Amplifier, Driver-Directional Listening	10A2	-	-	-	D	-		-	-	-	I	-	-		718435-801
✓	1	2	4	1			Capacitor	10A2 C15	-	-	-	-	-	-	-		-	D	-	-		CK06BX105K
✓	1	2	4	2			Relay, Armature	10A2 K1	-	-	-	-	-	-	-	-	-	D	-	-		8SBC1206A2
	1	2	4	3			Insulator, Relay		-	-	-	-	-	-	-	-	-	D		-		7717-129N
✓	1	2	4	4			Resistor	10A2 R1	-	-	-	-	-	-	-	-	-	D	-	-		RCR07G223J3
✓	1	2	4	5			Resistor	10A2 R2	-	-	-	-	-	-	-	-	-	D		-		RCR07G223J3
✓	1	2	4	6			Resistor	10A2 R3	-	-	-	-	-	-	-		-	D	-	-		RCR07G105J3
✓	1	2	4	7			Semiconductor Device	10A2 CR5	-	-	-	-	-	-	-	-	-	D	-	-		JAN1N645
✓	1	2	4	8			Resistor	10A2 R32	-	-	-	-	-	-	-	-	-	D	-			RN55D1783F
✓	1	2	4	9			Capacitor	10A2 C5	-	-	-	-	-	-	-	-	-	D	-	-		CM05FD221J03
✓	1	2	4	10			Transistor	10A2 Q4	-	-	-	-	-	-		-	-	D	-	-		JAN2N930
✓	1	2	4	11			Transistor	10A2 Q3	-	-	-	-	-	-	-	-	-	D	-	-		JAN2N930
	1	2	4	12			Insulator, Transistor		-	-	-	-	-		-	-	-	D	-	-		10197DAP
	1	2	4	13			Insulator, Transistor		-	-	-	-	-	-	-	-	-	D	-	-		10197DAP
✓	1	2	4	14			Capacitor	10A2 C13	-	-	-	-	-	-	-	-	-	D	-	-		CK06BX104K

Figure 4 - Example of a Task Identification Matrix (TIM). Cell entries: - (dash) no maintenance task of this type is performed on this hardware item; 0 - task of type, performed at organizational level; I - task, performed at intermediate level; and D - task, performed at depot level.

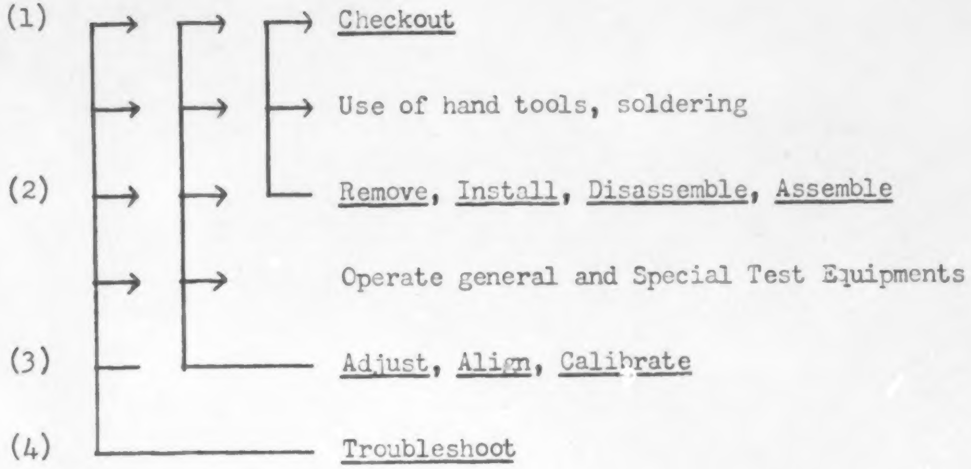

(1)
Checkout
Use of hand tools, soldering
(2)
Remove, Install, Disassemble, Assemble
Operate general and Special Test Equipments
(3)
Adjust, Align, Calibrate
(4)
Troubleshoot

Figure 5 - Indicating the Dependencies among Maintenance Functions for an Electronic Hardware (Functions Underlined)

makes it difficult to generalize from results of PM from one electronic hardware to another. One exception is in the area of general test equipment which may be used in performing maintenance tasks across many hardware subsystems.

The examples given are characteristic of many of the electronic maintenance AFSCs. Similar problems in complexity of maintenance functions and tasks are found in mechanical hardware, but to a lesser degree.

Development of PM and Symbolic Substitutes for PM

Starting in 1969, AS supported a modest program to provide the Air Force with the necessary tools for measuring the ability of maintenance personnel to perform the key tasks of their jobs. The scope of this work was limited to the maintenance of electronic hardware at the organizational and intermediate levels (see shaded portion of Figure 1). This program has two objectives: (1) to develop a model battery of JTPT together with appropriate scoring schemes for the measurement of the task performance ability of electronic maintenance personnel (an effort was to be made for the development of JTPT which could be easily administered), and (2) using the JTPT of this battery as criteria, to develop and try out a series of paper-and-pencil symbolic substitute tests that would hopefully have high empirical validity.

Criterion Referenced Job Task Performance Tests

A model battery of 48 criterion referenced JTPT and a test administrator's handbook were developed for measuring ability to perform electronic maintenance tasks. Copies of the actual instructions for test subjects together with the test administrator's handbook are available from the Defense Documentation Center (DDC) as AFHRL-TR-74-57(II) Part II (Shriver, Hayes, & Hufhand, 1975). The test administrator's handbook was developed with step-by-step detailed instructions so that an individual with a minimum of electronic maintenance experience can administer the tests.

The battery includes separate tests for the following classes of job activities: (1) equipment checkout, (2) alignment/calibration, (3) removal/replacement, (4) soldering, (5) use of general and special test equipment, and (6) troubleshooting. The Doppler Radar AN/APN-147 and its Computer AN/ASN-35 were selected as a typical electronic system. This system was used as the test-bed for this model battery. The soldering and general test equipment JTPT are applicable to all electronic technicians. The other tests of the battery apply to technicians concerned with this specific doppler radar system. A detailed description of the development and tryout of these JTPT is given in AFHRL-TR-74-57(II) Part I (Shriver & Foley, 1974a). Each class of activity for which JTPT was developed contains its individual mix of behaviors, but it is not mutually exclusive. As indicated in Figure 5 and Table 1, a four-level hierarchy of dependencies exists among them.

After considering product, process, and time as to their appropriateness for scoring the results for each activity, it was decided that a test subject had not reached criterion until he had produced a complete, satisfactory product. This was a go, no-go criterion.

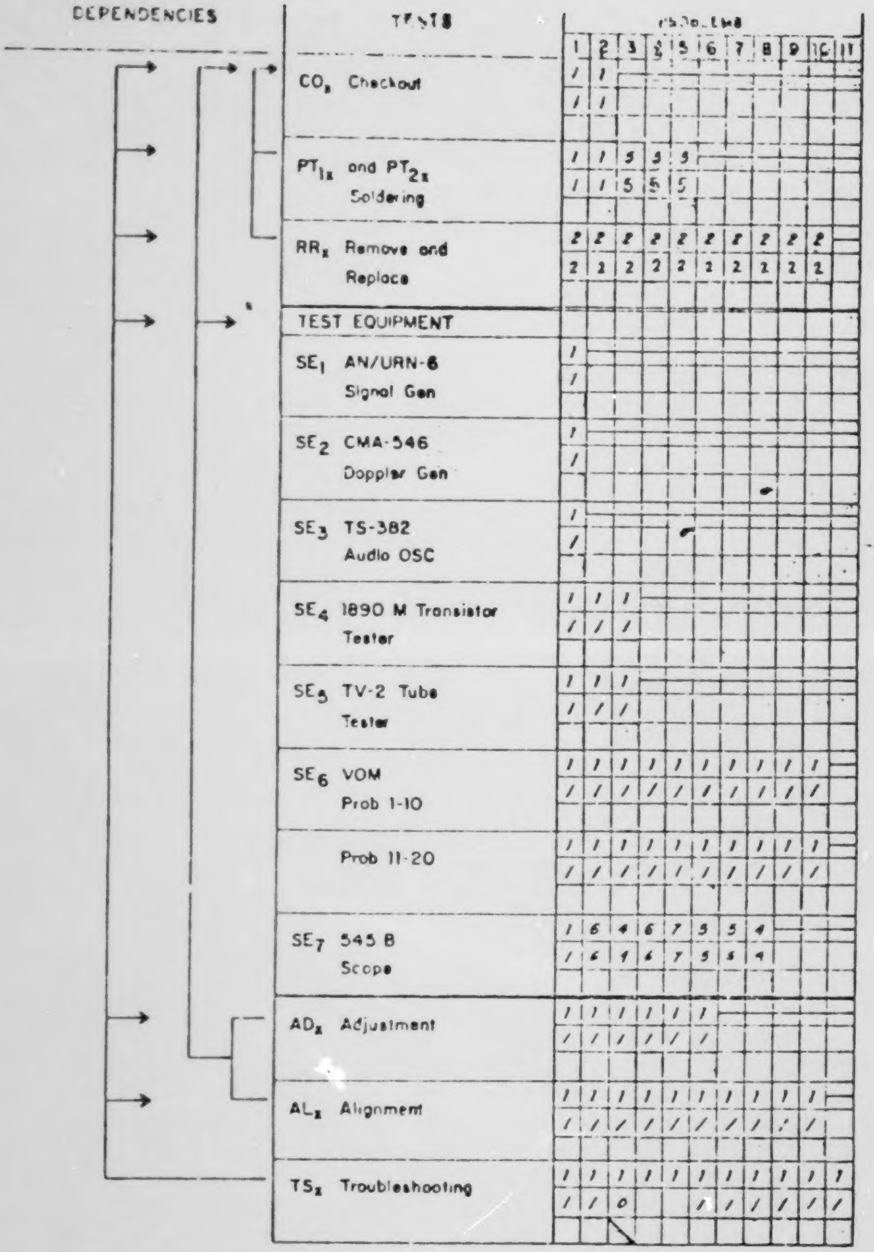
Table 2 summarizes the number of tests, problems, and scorable products by class developed for the AN/APN-147 and AN/ASN-35. The simple addition of numbers shown in Table 2 indicates that there are 48 tests, 81 problems, and 133 scorable products. But these numbers tell us nothing in terms of the content of the tests. To say that one test subject accomplished 100 scorable products while another accomplished 90 tells us nothing about the job readiness of these individuals or that one is better than the other. The varieties of scorable products are so diverse that any combination of them, without regard to what they represent, is meaningless. The only meaningful presentation of such information must be in terms of a profile designed to attach meaning to such numbers. A sample of such a profile is shown in Figure 6.

Table 2. Tests, Problems, and Scorable Products

Class	Code	Tests	Problems	Scorable Products
1. Checkout	CO	2	2	2
2. Physical Skills Tasks (soldering)	PT	2	5	17
3. Remove and Replace	RR	10	10	20
4. Test Equipment	SE	7	37	67
5. Adjustment	AD	6	6	6
6. Alignment	AL	10	10	10
7. Troubleshooting	TS	11	11	11
Total	7	48	81	133

This profile is not presented as the final solution to the profile problem for JTPT for electronic maintenance. It does contain most of the important information regarding a test subject's job task abilities as measured by the test battery, indicating the subject's strengths and weaknesses.

An examination of the profile (Figure 6) indicates that most of the tests in this battery contain only one problem. For example, there are two checkout tests having one problem each, and there are 11 troubleshooting tests having one problem each. There are two soldering tests; one has two problems and the other has three. The voltohmmeter (VOM) test has 20 problems.



TESTS	PROBLEMS										
	1	2	3	4	5	6	7	8	9	10	11
CO_x Checkout	/	/									
	/	/									
PT_{1x} and PT_{2x} Soldering	/	/	5	5	5						
	/	/	5	5	5						
RR_x Remove and Replace	2	2	2	2	2	2	2	2	2	2	
	2	2	2	2	2	2	2	2	2	2	
TEST EQUIPMENT											
SE_1 AN/URN-6 Signal Gen	/										
	/										
SE_2 CMA-546 Doppler Gen	/										
	/										
SE_3 TS-382 Audio OSC	/										
	/										
SE_4 1890 M Transistor Tester	/	/	/								
	/	/	/								
SE_5 TV-2 Tube Tester	/	/	/								
	/	/	/								
SE_6 VOM Prob 1-10	/	/	/	/	/	/	/	/	/	/	
	/	/	/	/	/	/	/	/	/	/	
Prob 11-20	/	/	/	/	/	/	/	/	/	/	
	/	/	/	/	/	/	/	/	/	/	
SE_7 545 B Scope	/	6	4	6	7	5	5	4			
	/	6	4	6	7	5	5	4			
AD_x Adjustment	/	/	/	/	/	/					
	/	/	/	/	/	/					
AL_x Alignment	/	/	/	/	/	/	/	/	/	/	
	/	/	/	/	/	/	/	/	/	/	
TS_x Troubleshooting	/	/	/	/	/	/	/	/	/	/	/
	/	/	0			/	/	/	/	/	/

Figure 6. A profile for displaying the results obtained by an individual subject from a battery of Job Task Performance Tests concerning an Electronic System – the AN/APN-147 and the AN/ASN-35. This represents the profile of an individual who has successfully completed most of the battery.

The subject receives no "credit" for a problem unless he obtains all of the expected products. No attempt is made to combine these scores in terms of meaningless numbers.

The hierarchy of dependencies discussed previously (Figure 5) has implications for the order in which tests are administered, as well as for diagnostics. For example, since troubleshooting includes the use of test equipment and other activities in the hierarchy, logic would dictate that in most training situations the administration of the tests for the sub-activities would precede the troubleshooting tests and that a test subject would not be permitted to take the troubleshooting tests until he had passed these other subtests. Under some circumstances, one may wish to reverse the process. A subject who successfully completes selected troubleshooting or alignment tests can be assumed to be proficient in his use of test equipment and checkout procedures. These dependencies are displayed on the left-hand side of the profile (Figure 6).

Due to the unavailability of a sufficient number of experienced test subjects at the time of the tryout of the JTPT battery, the tryout was not as extensive as planned. The limited tryout did indicate that the tests as developed are administratively feasible. Their continued use, no doubt, would result in further modifications and improvements.

Development of Symbolic Substitutes

There is no doubt that a battery of JTPT would require more training and on-the-job time of the test subjects, more equipment, and specially trained test administrators. Therefore, the availability of empirically valid symbolic substitute tests would be highly desirable. Even though previous attempts to develop such tests as the Tab test (Crowder, Morrison, & Demaree, 1954) had failed, it was our opinion that much more work could be done to improve symbolic maintenance tests as substitutes for JTPT. It was hypothesized that higher correlations possibly could be obtained by a different approach to the development of symbolic tests. A study of the Tab Tests (Crowder et al., 1954, see Table 1) indicated that the JTPT used as the criterion measures contained many distractions and interruptions to the subject's troubleshooting strategy (cognitive process); such as using test equipment to obtain test point information. In addition to such interruptions to the cognitive process, the subject can obtain faulty test point information by the improper use of his test equipment. In the symbolic substitute Tab Tests, all of these potential pitfalls of the actual task were avoided. The subject was given a printed test point readout. It was hypothesized that the injection of job equivalent pitfalls into symbolic substitutes possibly would increase their empirical validity.

Based on these hypotheses, a battery of symbolic tests were developed under contract with the Matrix Research Company of Falls Church, Virginia.

A companion graphic symbolic test was developed for each of the job activities for which a criterion referenced JTPT had previously been developed. Based on two limited validations, all of the graphic symbolic tests, with the exception of the symbolic test for soldering, indicated sufficient promise to justify further consideration and refinement. Table 3 indicates the correlations obtained from these validations. Due to a shortage of available subjects, the number of pairs of subjects was extremely small. All of these promising graphic symbolic tests, therefore, must be given more extensive validations using larger numbers of experienced subjects.

The validation of any such symbolic test requires the administration of a companion JTPT as a validation criterion. As a result, a validation is an expensive process in terms of equipment and experienced manpower. The troubleshooting symbolic tests require the most extensive refinement. Several suggestions are made for improving their empirical validity. A complete description of these symbolic test efforts can be found in AFHRL-TR-74-57(III) (Shriver & Foley, 1974b). An attempt, also, was made to develop video symbolic substitute tests, but this effort produced no promising results (Shriver & Hufhand, 1974).

Even if graphic symbolic substitutes of high empirical validity can be produced, the use of symbolic substitutes will never, in my opinion, dispense with the requirement for the liberal administration of actual JTPT to maintenance personnel. We can never include all aspects of the actual performance of a task in a paper-and-pencil symbolic representation of that task, but our work indicates that we can come much closer than has been done in the past.

The Sampling Problem

Timewise, it would be impossible to administer a JTPT to a maintenance man for every possible task that his hardware system might produce. This world of tasks and people must be sampled. The model battery described previously provides a sampling procedure based on major task functions such as checkout, align, adjust, troubleshoot, etc. But even this sampling across possible tasks resulted in 48 tests and 133 scorable products (Table 2). It would be impractical to give any one test subject all these 48 tests at any one time. Systematic sampling schemes must be developed across tests.

The purpose for which JTPT results are to be used should be considered when developing sampling schemes. Such purposes could include ascertaining (1) the job task proficiency of an individual, (2) the job effectiveness of a training program, and (3) the proficiency of a maintenance unit. Each of these purposes would require a different mix or mixes of tests and people. Some suggestions for such samplings can be found in AFHRL-TR-74-57(II) Part I (Shriver & Foley, 1974a). But it should be remembered that these are suggestions that must still be field tested.

In the case of determining unit proficiency, some JTPT can be administered by on-line observation of tasks which are often repeated such as checkout. There will always be a requirement for off-line PM concerning critical, but seldom performed tasks. Whether the JTPT is performed on-line or off-line, the test administrator must use the same objective scoring procedures, the criteria of success being an acceptable product.

Consolidated Data Base to Support PM

In keeping with its man-machine interface orientation, AFHRL/AS is demonstrating the technical feasibility of integrating five human resources related technologies and applying them during weapon system development. This is being accomplished under Project 1959, "Advanced System for the Human Resources Support of Weapon System Development."

The five technologies are:

Human Resources in Design Tradeoffs
Maintenance Manpower Modeling
Job Performance Aids
Instructional System Design
System Ownership Costing

One objective of this program is to determine the data input requirements for and prepare specifications for a consolidated maintenance task identification and analysis data base which will support the integrated application of these five technologies in a weapon system development program. We feel that such a consolidated data base will contain most, if not all, of the information which would be required to develop good JTPT provided the tests are developed in keeping with the technology described in this paper. If such a data base is demonstrated to be technically feasible and if it is routinely made a requirement in weapon system development contracts, it will provide considerable assistance in developing maintenance performance tests for new weapon systems.

Institutionalization of New Technologies

Getting newly developed technologies such as PM institutionalized is a perennial problem, especially when a technology requires fundamental changes in long existing programs, procedures, and attitudes of entrenched establishments. AS has been involved in the implementation of several well developed and documented technologies, such as job performance aids and instructional system design (ISD) including programmed instruction and job (task) oriented training. These experiences have indicated that it is extremely difficult to maintain the integrity of a technology during its so called implementation. Operational organizations invariably attempt to implement a much "watered down" version of the technology and consequently obtain much "watered down" results. In some cases, only cosmetic changes to existing programs are reported as implementations. Currently, it requires many years of persistent effort on the part of the research community to get a technology properly institutionalized.

A mechanism must be developed for the timely institutionalization of each new technology which will ensure its integrity. A mechanism for the orderly implementation of technologies, similar to that used for new weapons systems, is recommended. Such a mechanism must make efficient and effective use of the "know-how" of the developers of the technology and make them responsible and accountable for its implementation. A new technology should not be turned over to a using command for its operation until it is in place, "debugged" and operational--just as a new weapons system is not turned over to an operational command until it has been "debugged" and proven to be ready for operational use.

Proposed PM R&D Efforts for Maintenance

Excessive maintenance costs are never going to be reduced as long as we don't have JTPT and/or empirically valid symbolic substitutes to ascertain how efficiently maintenance men perform the tasks of their jobs. In my opinion, the lack of such measures of maintenance performance is a most serious deficiency in DoD. As such, R&D in this area should have an extremely high priority.

Areas for R&D Concentration

For a long-range R&D effort, five general areas of concentration are recommended: namely, JTPT and matching symbolic substitute tests for electronic maintenance, JTPT and matching symbolic substitute tests for mechanical maintenance, and aptitude tests based on PM. The development and field tryout of a JTPT must precede the development of its symbolic substitute. The work on JTPT batteries for both electronic and mechanical maintenance should be started as soon as possible. The work on aptitude tests should not be started until JTPT batteries and the symbolic substitute tests have been completely field tested. More information concerning these areas of concentration follows:

1. Refinement of Model JTPT Battery (Electronic Maintenance)--The already available model JTPT Battery (Shriver, Hayes, & Hufhand, 1975) should be given a large scale field tryout. (Since the AB328X4 Avionics Inertial and Radar Navigation Systems Specialist Course, which includes the AN/APN-147 and the AN/ASN-35, does not emphasize the mastery of job tasks, the equipment specific tests of this battery cannot be used in the formal course.) One thrust of this effort should be to further refine the battery including its administrative procedures. A second thrust should be the development of sampling strategies which would be appropriate for determining the effectiveness of training programs and both individual and unit proficiency as discussed earlier under PM

102

problems. This effort would require approximately 2 professional man-years plus the use of maintenance specialists as test administrators from the appropriate maintenance specialties. If it is necessary to select a system other than the AN/APN-147-AN/AJN-35 combination, this work would require approximately 4 professional man-years.

2. Refinement of Symbolic Substitutes (Electronic Maintenance)--As previously indicated, a number of symbolic substitutes for JTPT were developed and given a limited tryout. Table 3 indicated that some of the symbolic tests show promising empirical validity. These promising symbolic tests must be more thoroughly refined and validated. In addition, further exploratory development is required for symbolic substitute tests for troubleshooting tasks in keeping with recommendations made in AFHRL-TR-74-57(III) (Shriver & Foley, 1974b). This effort would require between 3 and 4 professional man-years plus the use of maintenance specialists as test administrators and test subjects from the appropriate maintenance specialties.

Table 3. Indicates the Numbers of Pairs Used as Well as the χ^2 and the Correlations Obtained During Two Small Validations of Symbolic Tests

Test Areas	N Pairs	χ^2	ϕ	r_t
Novice Subjects (Altus)				
Checkout	4	4.00	1.00	-
Remove & Replace	14	2.57	.43	-
Soldering Tests	4	0	0	-
General Test Equip	6	2.67	.67	-
Special Test Equip	6	.67	.33	-
Alignment/Adjustment	19	6.37	.58	-
Troubleshooting	9	1.00	-.33[a]	-
Experienced Subjects (TAC)				
Overall Troubleshooting	30	6.53	.47	.68
Chassis (Black box)				
Isolation	30	16.33	.73	.81
Stage Isolation	30	3.33	.33	.46
Piece/Part Isolation	15	.07	.07	.16

[a] This negative correlation was probably due to a number of deficiencies such as (1) deficiencies in the Fully Proceduralized Job Performance Aids provided the subjects, (2) deficiencies in the sequencing of the troubleshooting JTPT in relation to the sub-tests in the JTPT battery, (3) maintenance difficulties with the AN/APN-147 AN/ASN-35 system, and (4) difficulties with the content and administration of test equipment pictorials provided in the original troubleshooting symbolic tests.

3. Development of Model JTPT Battery (Mechanical Maintenance)--A model JTPT battery similar to the model battery for electronic maintenance described previously should be developed for a typical mechanical subsystem such as a jet engine or tank engine covering both the organizational and intermediate levels of maintenance. This model should be thoroughly field tested. Sampling strategies as indicated for the electronic battery should also be developed. This effort will require approximately 4 professional man-years plus the use of maintenance men from the appropriate maintenance specialties as test administrators and test subjects.

4. Development of Symbolic Substitutes (Mechanical Maintenance)--An attempt should be made to develop symbolic substitute tests with high empirical validity after the model JTPT battery is available for mechanical maintenance. The same contractor should develop these symbolics as developed the JTPT battery. A very rough estimate for accomplishing this symbolic effort would be 4 professional man-years.

5. Job Aptitude Test Research Based on Results on JTPT--R&D plans should be made to utilize the results of JTPT and symbolic substitute tests for standardizing military aptitude indices obtained from the Armed Services Vocational Aptitude Battery (ASVAB). As a first step, the military aptitude scores of all test subjects used for the tryouts in the proposed JTPT R&D should be recorded. In addition, such aptitude scores should be obtained during any school or field administration of JTPT or symbolic substitutes. When sufficient data are obtained, the degree of relationship between JTPT results and various aptitude indices should be obtained. Later, when a sufficient number of JTPT are used in the field, a formal R&D project should be initiated to modify the ASVAB to directly reflect job success as measured by JTPT.

R&D Strategy

Probably the most cost-effective approach for PM for both electronic and mechanical maintenance would be to concentrate on the development and refinement of JTPT on use of key test equipments prior to proceeding with the other task functions of the proposed model test batteries. As indicated in Figure 5, the use of general test equipment is a prerequisite to maintenance task functions such as alignment, calibration, and troubleshooting. In addition, general test equipments usually have wide usage in such task functions across many hardware systems, and there is a substantial amount of data which indicates that many maintenance men are weak in their test equipment ability. So, a general improvement in ability to use test equipment is an important and necessary factor for the general improvement of several maintenance task functions. I would strongly recommend, therefore, the early concentration for the proposed model test batteries in this area. Each PM development for a test equipment should be accompanied by the development of a programmed training package with sufficient practice frames for teaching the mastery of all its functions. Basic models of such training packages for 12 general test equipments are not available (see Scott & Joyce, 1975a through 19751). However, more practice frames should be included in these programs.

Closing Statement

Maintenance of hardware is currently an extremely costly operation for the DoD. High maintenance cost is the primary cause of high systems ownership cost. For some electronic maintenance specialties, nearly 1 year of broad formal training is given first enlistment personnel. And maintenance training generally is long and costly. Even with such lengthy training, the efficiency of maintenance could be greatly improved. Improved job instructions and information as well as increased use of job (task) oriented training have great potential for decreasing maintenance training time and improving the job performance of maintenance tasks. But to realize such potential, the criteria for the personnel system (selection, training, assignment, and promotion) for maintenance personnel must be shifted to the demonstrated ability to perform the tasks of their jobs. (The current criteria emphasize the ability to obtain high scores on paper-and-pencil theory and job knowledge tests.)

In this paper, I have discussed what I think are the important aspects of the criterion problem as it applies to the measurement of ability to perform maintenance tasks in training and on-the-job. Our objective in its solution is to get as close to the real job as possible. When "on-line" tasks occur often enough, their structured observation may be appropriate. But when such observations are not appropriate or when tasks occur infrequently, we propose to have the tasks performed "off-line" in a job-like environment. Our approach to the development of such measures was started with an analysis of the structure maintenance of the man/hardware interface. Based on the results of this analysis, we developed a model test battery of JTPT for electronic maintenance. Using this model as the criterion, we also developed batteries of graphic and video symbolic substitute tests. Several of the graphic symbolics have indicated respectable empirical validities but require more refinement and tryout. Our attempts to develop video symbolics were unsuccessful.

I have recommended a research program based on what we have already accomplished. This includes the development of a model battery of JTPT together with symbolic substitutes for maintenance tasks generated by a typical mechanical hardware. I have also discussed briefly the perennial problems of getting new technologies such as JTPT implemented. There is definitely a requirement for a structured mechanism which will guarantee the orderly institutionalization of such technologies as well as their integrity during the implementation process.

REFERENCES

Anderson, A.V. Training, utilization, and proficiency of Navy electronics technicians: III, Proficiency in the use of test equipment (Navy Technical Bulletin 62-14, AD-294 330). San Diego: U.S. Navy Personnel Research Activity, 1962.

Brown, G.H., Zaynor, W.C., Bernstein, A.H., & Shoemaker, H.A. Development and evaluation of an improved field radio repair course (Technical Report 58, Project Repair, AD-227 173). Washington DC: Human Resources Research Office, The George Washington University, 1959.

Crowder, N., Morrison, E.J., & Demaree, R.G. Proficiency of Q-24 Radar mechanics: VI. Analysis of intercorrelations of measures (AFPTRC-TR-54-127, AD-62 116). Lackland AFB TX: Air Force Personnel and Training Research Center, 1954.

Evans, R.N., & Smith, L.J. A study of performance measures of troubleshooting ability on electronic equipment. Illinois: College of Education, University of Illinois, October 1953.

Foley, J.P., Jr. Critical evaluation of measurement practices in post-high school vocational electronic technology courses (AD-683 729). Doctoral dissertation, University of Cincinnati, 1967.

Foley, J.P., Jr. Description and results of the Air Force research and development program for the improvement of maintenance efficiency (AFHRL-TR-72-72, AD-77 100). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, November 1973.

Foley, J.P., Jr. Evaluating maintenance performance: An analysis (AFHRL-TR-74-57(I), AD-A004 761). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, October 1974.

Foley, J.P., Jr. Factors to consider in developing new test and evaluation techniques. Proceedings of the Human Factors Testing Conference, 1-2 October 1968. Snyder, M.T. (Chm.), Kincaid, J.P., & Potempa, K.W. (Eds.), AFHRL-TR-69-6, AD-866 485. Wright-Patterson AFB OH; Advanced Systems Division, Air Force Human Resources Laboratory, October 1969.

Frederiksen, N. Proficiency tests for training evaluation. In R. Glaser (Ed.), Training Research and Education. Pittsburgh PA: University of Pittsburgh Press, 1962.

Harris, D., & Mackie, R.R. Factors in influencing the use of practical performance tests in the Navy (Navy Technical Report No. 703-1, AD-284 842). Washington DC: Office of Naval Research, 1962.

Jenkins, J. G. Validity for what? Journal of Consulting Psychology, March-April 1946.

Joyce, R.P., Chenzoff, A.P., Mulligan, J.F., & Mallory, W.J. Fully proceduralized job performance aids: Draft military specification for organization and intermediate maintenance (AFHRL-TR-73-43(I), AD-775 702). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, December 1973.

Mackie, R.R., Wilson, C.L., & Buckner, D.N. Practical performance test batteries for electricians mates and radiomen developed in conjunction with a manual for use in the preparation and administration of practical performance tests (AD-98 239). Los Angeles CA: Management and Marketing Research Corporation, June 1953.

Saupe, J.L. An analysis of troubleshooting behavior of radio mechanic trainees (AFPTRC-TN-55-47, AD-99 361). Lackland AFB TX: Air Force Personnel and Training Center, November 1955.

Scott, D.L., & Joyce, R.P. TEKTRONIX 545B oscilloscope training (AFHRL-TR-76-19, AD-A022 941). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (a)

Scott, D.L., & Joyce, R.P. TS-1100/U transistor tester training (AFHRL-TR-76-20, AD-A022 930). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (b)

Scott, D.L., & Joyce, R.P. TS-148 radar test set training (AFHRL-TR-76-21, AD-A022 931). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (c)

Scott, D.L., & Joyce, R.P. TV-2a/U tube tester training (AFHRL-TR-76-22, AD-A-22 932). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (d)

Scott, D.L., & Joyce, R.P. URM-25D signal generator training (AFHRL-TR-76-23, AD-A022 933). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (e)

Scott, D.L., & Joyce, R.P. 200 CD wide range oscillator training (AFHRL-TR-76-24, AD-AD-A022 934). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (f)

Scott, D.L., & Joyce, R.P. 5245 L electronic counter training (AFHRL-TR-76-25, AD-A022 939). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (g)

Scott, D.L., & Joyce, R.P. *Fluke 803 differential voltmeter training* (AFHRL-TR-76-26, AD-A022 956). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (h)

Scott, D.L., & Joyce, R.P. *HP-410B VTVM training* (AFHRL-TR-76-27, AD-A022 940). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (i)

Scott, D.L., & Joyce, R.P. *Kay model 860 sweep generator training* (AFHRL-TR-76-28, AD-A022 957). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (j)

Scott, D.L., & Joyce, R.P. *SG-299 B/U signal generator training* (AFHRL-TR-76-29, AD-A022 972). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (k)

Scott, D.L., & Joyce, R.P. *Simpson 260 VOM training* (AFHRL-TR-76-30, AD-A022 984). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (l)

Shriver E.L., & Foley, J.P., Jr. *Evaluating maintenance performance: The development and tryout of criterion referenced job task performance tests for electronic maintenance* (AFHRL-TR-74-57(II), Part I, AD-A004 845). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1974. (a)

Shriver, E.L., & Foley, J.P., Jr. *Evaluating Maintenance Performance: The development of graphic symbolic substitutes for criterion referenced job task performance tests for electronic maintenance* (AFHRL-TR-74-57 (III), AD-A005 296). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, November 1974. (b)

Shriver, E.L., & Foley, J.P., Jr. *Job performance aids for UH-1H helicopter: Controlled field tryout and evaluation* (AFHRL-TR-75-28(I), AD-B006 295L. Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, June 1975. (Distribution limited to U.S. Gov't. agencies only.)

Shriver, E.L., Hayes, J.F., & Hufhand, W.R. *Evaluating maintenance performance: Test administrator's manual and test subject's instructions for criterion referenced job task performance tests for electronic maintenance* (AFHRL-TR-74-47(II), Part II, AD-A005 785). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, January 1975.

Shriver, E.L., Hayes, J.F., & Hufhand, W.R. *Evaluating maintenance performance: A video approach to symbolic testing of electronics maintenance tasks* (AFHRL-TR-74-57(IV), AD-A005 297). Wright-Patterson AFB OH: Advanced Systems Division, Air Force Human Resources Laboratory, July 1974.

Vineberg, R., Taylor, E.N., & Caylor, J.S. *Performance in five Army jobs by men at different aptitude (AFQT) levels: 1. Purpose and design of study* (TR-70-18, AD-715 614). Presidio of Monterey CA: Human Resources Research Organization, 1970. (a)

Vineberg, R., Taylor, E.N., & Sticht, T.G. *Performance in five Army jobs by men at different aptitude (AFQT) levels: 2. Development and description of instruments* (TR-70-20, AD-720 216). Presidio of Monterey CA: Human Resources Research Organization, 1970. (b)

Vineberg, R., & Taylor, E.N. *Performance in four Army jobs at different aptitude (AFQT) levels: 3. The relationship of AFQT and job experience to job performance* (TR-72-22, AD-750 360). Presidio of Monterey CA: Human Resources Research Organization, 1972. (a)

Vineberg, R., & Taylor, E.N. *Performance in four Army jobs at different aptitude (AFQT) levels: 4. Relationships between performance criteria* (TR-72-23, AD-750 604). Presidio of Monterey CA: Human Resources Research Organization, 1972. (b)

Wallace, S.R. Criteria for what? *American Psychologist*, June 1965. (a)

Wallace, S.R. The relationship of psychological evaluation to needs of the Department of Defense. *Proceedings of 7th Annual Military Testing Association Conference* (AD-681-096). San Antonio TX: October 1965b, 1-10. (b)

Williams, W.L., Jr., & Whitmore, P.G., Jr. *The development and use of a performance test as a basis for comparing technicians with and without field experience* (The NIKE AJAX AFC Maintenance Technician Technical Report 52, AD-212 663). Washington DC: Human Resources Research Office, The George Washington University, January 1959.

FOOTNOTES

Extraneous remarks by Dr. Foley

1. I want to say something here. I said, "for reducing training time." I want to make it clear that I didn't say "reducing training cost," because I've been accused of that. Your training costs, when you get into job oriented training, go up--or at least stay the same--your training costs per course are probably about the same. The only thing is they're more costly per week, but by reducing training time you do reduce cost as time in the field, for the more time you have a man in the field in his first enlistment, the less often you have to replace him.

2. Now, we don't have quite that bad a situation, but we cover up that situation in the field of maintenance by gobbling up a lot of spare parts, and that's been costing us all kinds of money. Anytime we can get our hands on spare parts that have been turned in, we find that a great many of them are still good but they are destroyed because people are what we call "shot-gunned" and found a faulty part by removing and replacing a large number of good parts.

XI

CRITERION PROBLEMS

Cecil J. Mullins and Forrest R. Ratliff
Personnel Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

[1]When we first began struggling in this wonderfully complex area of criterion analysis and development, we were almost overwhelmed by the assortment of special and seemingly divergent problems associated with criterion variables. These were problems that seemed to be unrelated to predictor research, and even unrelated to each other. For instance, how ultimate should a criterion be? Are we trying to select people who will do well in training, or those who will perform satisfactorily on their first job, or those who will get through their first hitch, or those . . .? Previous work (Ghiselli & Haire, 1960; Prien, 1966) shows rather clearly that those subjects who are high on some proximal standard are not necessarily high on any of the more distal ones.

Also, what is the best way to collect criterion information? Ratings are cheap and they have a certain ring of truth to the rater, but we know that ratings rarely work well, particularly in the operational situation. Assessment centers and job-sample data are far too expensive for routine evaluation of subjects, and there are certain conceptual difficulties even with them. How does one collect performance data in one situation in such a way that the scores issuing from the exercise are comparable with scores on other people doing essentially the same work but in a different condition, with a different supervisor, and a different social climate? How does one even demonstrate that a particular criterion variable is good or bad? Somehow, it jars to talk about "validating" a criterion.

All in all, the most serious difficulty we had was the lack of a philosophy or orientation. We needed some way of organizing our approach, some framework which might systematize our thought and our efforts. We have come to a way of thinking about the problem which, at least for us, has proved somewhat helpful.

Let us consider what we mean by the word, "criterion." Of course, there is the purely statistical meaning of the term, which means simply a target variable which we are trying to reproduce by appropriate mathematical manipulation of other variables. Statistically, the criterion could be any variable, and the predictor could be any other variable. But I am referring to the conceptual meaning of "criterion," as distinct from the word "predictor." Let us examine some of the faulty ideas I have held for several years about criterion and predictor variables. I don't know if anyone here has ever held these ideas, but I do find them rather widespread. We are not talking here about formal definitions, but only about conventional wisdom.

Example 1. Predictors are aptitude-type variables and criteria are achievement measures revealed by some kind of performance. I grew up with this idea, and I have since found it to be a fairly common misconception. Actually, practically all psychometric variables are achievement measures. We are not by any means the first to notice this (Thorndike, 1926; Estes, 1974). Tests of verbal aptitude, for example, are usually tests of a subject's current achieved ability to perform with words. All aptitude measures that I can think of are really tests of achievement, just like criterion tests. On the other hand, it is generally accepted that the best predictor of future achievement is past achievement. Upon examination, then, this distinction between criterion and predictor disappears.

Example 2. Predictor variables usually represent something "basic"--perhaps even genetic--while criterion measures represent some sort of ultimate achievement acquired by the subject through training or experience. This distinction may be partially true, in that development of characteristics continues from birth to death. But we think it is not true in the sense in which it is frequently understood. To use verbal aptitude as an example again, there is no substantial evidence for the existence of verbal aptitude as a basic dimension of human ability except that it appears in one particular kind of factor analysis, and even then only if the data are collected on subjects older than a certain age. We think it likely that there are basic aptitudinal underlayments, probably genetic, but that these are far more simple and fundamental than the Thurstonian aptitudes. There are probably some very raw individual differences present at birth, similar to Horn's anlage functions (1968) or Cattell's fluid intelligence (1941).

To let Horn speak for himself:

> (The Anlage function) represents very elementary capacities in perception, retention and expression, as these govern intellectual performance. For example, span of apprehension--the number of distinct elements which a person can maintain in immediate awareness--is an elementary capacity and yet one which determines, in part, the complexity with which one can successfully cope in an intellectual task. It would seem that such capacities are not much affected by learning--anlage functioning is closely associated with neural-physiological structure and process--but that such functions operate to some extent

> in all intellectual performances and thus produce variance in all ability measurements.

Exactly what the anlage functions are, or even how many of them there are, is still a matter to be determined by research. Whatever they may be, they are seen as immutable individual differences--probably genetic--which remain stable and constant throughout the life of the individual. As we shall see later, the anlage functions can become overlaid by a considerable depth of learned material, so that their observation as pure characteristics is very difficult, but they exist nonetheless, in about the same quantities as they existed at birth.

It is only after certain other measurable conditions have occurred and have interacted that something as advanced as verbal (or numerical, or spatial) aptitude develops to a measurable degree. Thus, it is entirely logical that in some situations a test of verbal aptitude might be used as a criterion measure to be predicted by the more basic anlage functions. Similarly, later developments (say, performance in Psychology 201) might with equal logic constitute a criterion to be predicted by a verbal aptitude test, and some other behavior (say, progress as a research psychologist) may be predicted by grades in Psychology 201. In sum, then, there is nothing ultimate about any "criterion," and nothing basic about any "predictor" with the possible exception of those unknown anlage functions we just mentioned.

Example 3. Predictor variables are simple, factorially pure measures, and criteria are complex. Since development normally proceeds from more simple to more complex, and since criterion measures are usually taken later than predictor measures, this is probably true in a general sense. However, there is nothing absolute about this principle, either. For example, the last time I looked, the best single predictor of college performance (a criterion) was high school performance (a predictor). There are other, much purer, predictor measures, but they don't ordinarily do as good a job as the much more complex variable of high school grades.

Example 4. Predictor data are collected at an earlier time than criterion data. So far as we can tell, this is the only general statement one can accurately make about the distinction between criteria and predictors. All the other distinctions, as we have seen, either disappear entirely upon examination, or exist only partially and only some of the time.

So where does all this lead us? It seems to me to lead to the conclusion that there is no such thing as a "criterion" problem, distinct from "predictor" problems. There are only measurement problems, equally applicable to all measurement, whether predictors or criteria. The measurement problems concern the best ways to collect current status data, whether we call the data predictors or criteria, at various points of a subject's career.

When we speak of current status data, we are talking about achievement or, more precisely, intellectual development. We believe that intellectual development proceeds in some exponential manner, so that learning is built on learning according to some interaction among four general terms; previous learning, potential, opportunity, and energy. Of course, we don't yet know the exact formulation of the postulated relationship, but we feel that it should be something like $D_2 = D_1 \ (1 + i)^t$, where D_2 means development at some later time, D_1 means development at some earlier time, and "t" refers to units of time separating the two developmental points. The term "i" is a deceptively simple-looking term, which is anything but simple. It refers to some interaction of potential (the anlage functions), opportunity (measurable in a very crude degree by experience and training), and an energizing function (both physical and psychological, including interest, motivation, and similar concepts). This formula produces a constantly accelerating curve like those shown in Figure 1. Obviously, this is not yet a very practical working formula--there are too many unknowns in the terms--but it does have some use to us in helping us order our thinking. For example, this formula tells us that two people with different potential can arrive at the same state of development at the same time because of differences in opportunity and energy (lines A and C, Figure 1, converge between t_5 and t_6). Our practical experience tells us that, indeed, this sort of convergence does occur. Also, this orientation suggests that the best predictor of some developmental point (a criterion) is the nearest practical earlier point, measured fully. Otherwise, one must know much more than one usually knows about opportunity and energy, since the longer the time period separating the two points, the larger "t" becomes in the equation, and the more important opportunity and energy become. It has helped us a great deal in thinking about intellectual development, and criteria are, as we see them, only points on the curve of intellectual development.

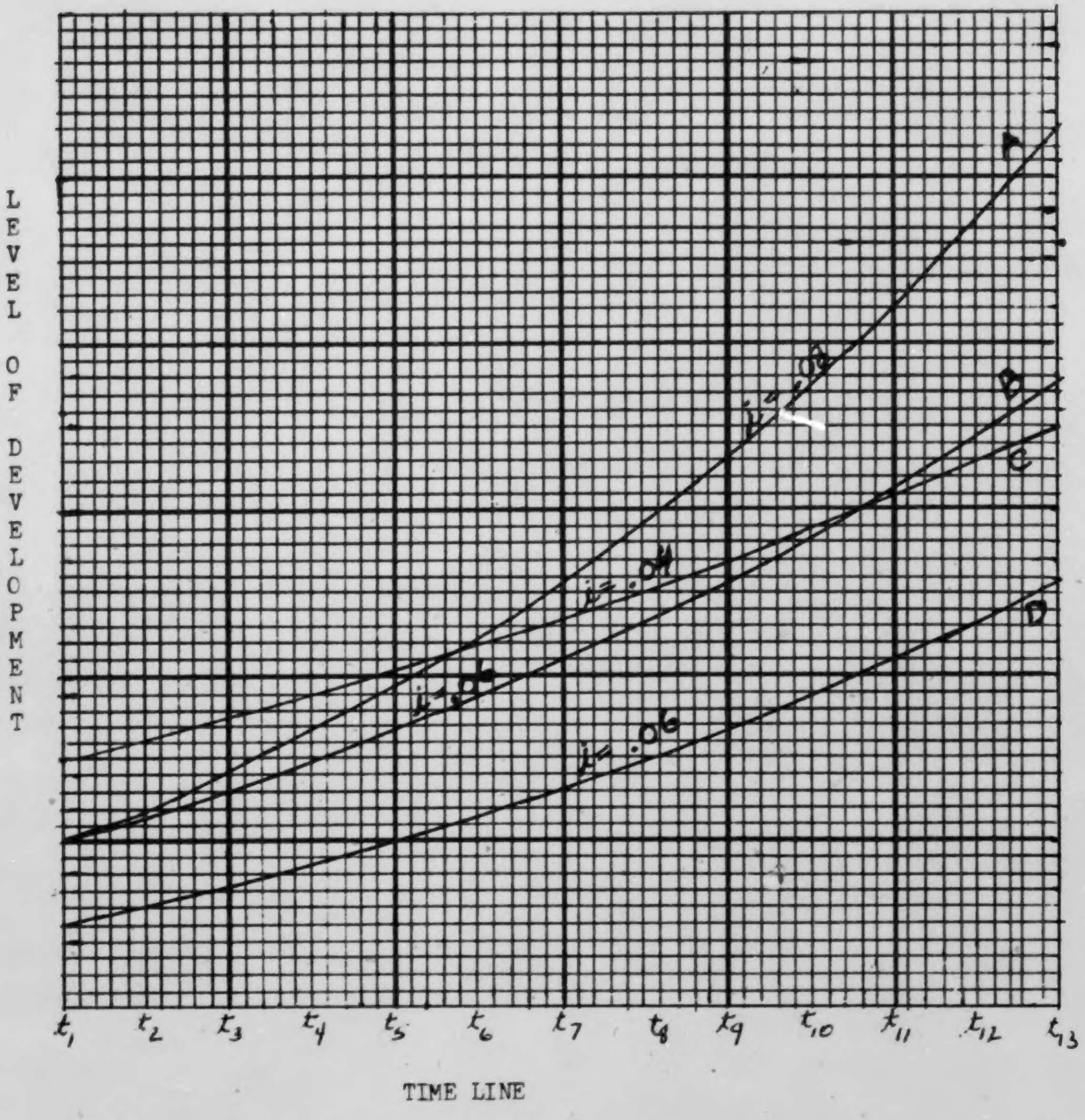
We have said there is no specific criterion problem--only measurement problems. Heaven knows these problems are severe enough. As we look at them, they fall into several dimensions. Keep in mind that all subject assessment, whether taken earlier as predictor information or later as criterion data, can be collected in the same ways and are afflicted by the same difficulties. There are no special difficulties unique to either predictors or criteria.

Kind of Data. Measurement data can be collected in many ways. Some of the most important ways are:

1. Ratings. We can ask the subject or someone else to give us an opinion. On the relatively low level of measuring aptitudes, we have

Figure 1. Generalized Lines of Intellectual Development

$D2 = D_1 (1 + i)^t$ $\qquad$ $i = \text{Pot} \longleftrightarrow \text{Opp} \longleftrightarrow \text{En}$



been able a long time back to move from opinions to tested performance. One reason for our success in that area has undoubtedly been our ability to validate and refine aptitude test ideas against various criteria. But we have not been so successful in this respect in our development of criterion measurement ideas. Possibly one reason for our lack of success here has been that we have not seen criteria for what they are--points along a development continuum followed by other points against which it should be possible to validate them. When we look at criteria in this way, it seems to me that we don't have to settle for the desperate position of Nagel (1953), Brogden and Taylor (1950), and others that criterion measures by their nature are always judgmental (i.e., not subject to verification). We can validate criteria against later criteria and proceed with criterion development in much the same way we have done with predictors. When we have brought the state-of-the-art a little higher, we can perhaps dispense with ratings as criterion data, just as we have done on the predictor side.

We shall see later that there is another, probably more important, reason that we use ratings so often as criteria. At any rate, ratings are now used much more often to collect criterion data than to collect predictor data. There are a few things to recommend ratings--they are quick and cheap and, under the right conditions, they can be made to yield useful information about the ratee. On the negative side, some problems inherent in the nature of rating data loom very large.

There appear to be individual differences (as one should suspect) in the ability of people to assess other people accurately. We are doing work on this phenomenon, which Mr. Weeks will tell you about later. Furthermore, even good raters are often put in a situation which militates against the collection of good information. If the ratee is to have access to the rating and if the rating is to influence the ratee's career in any way, it is not likely that a supervisor will produce ratings of his people which can be considered a good assessment tool. The supervisor is placed in a position which requires him to perform two mutually incompatible acts. As a supervisor, he is responsible directly or indirectly for the morale and energy of his work unit, which calls for support by him of his people; but he is also required to render an objective and accurate appraisal which is likely to damage some or all of those same subordinates. It is a rare supervisor who can do both. As a result, all the operational rating systems that I am aware of suffer the usual inflation of means and compression of variance. I do not believe there is any way that a useful criterion can be collected in the military environment from supervisor ratings collected operationally in the usual way, so we have to look for innovation. We are doing work which we think will alleviate this problem somewhat, and I shall report more fully on this effort later.

2. Job-sample tests. Job-sample tests, in their usual format, are prohibitively expensive for operational use. I say this despite the comments of several astute observers (e.g., Otis, 1953) who have

pointed out, in effect, that since good criterion information is absolutely basic to all personnel actions, we should consider any expense connected with its collection a very good investment. We believe that a certain amount of actual job simulation, or assessment center type evaluation, must be available for research purposes, but it is probably impractical to consider this kind of criterion for anything other than experimentation or in the evaluation of less expensive methods. We are embarking on an effort to capture as much of the essence of a job as possible on motion picture film, which can then be used as a test stimulus for collecting criterion information in large groups, thereby reducing its cost appreciably.

3. There are, of course, other ways to collect criterion information (e.g., paper-and-pencil tests), all of which pose problems which eventually we shall have to address. Some of the work we are doing is on paper-and-pencil criterion tests, the items of which are selected to maximize differences between subjects at different career levels. But regardless of how the data are collected, there are other dimensions of problems which must be considered also, so we must move on.

Use of Data.

Criterion data can be collected for many purposes--to promote, to serve as a target variable for predictor tests, to indicate need for training, to be used in reassignment of duties, and many more. When we consider a particular set of criterion data, we should clarify as early as possible what use is to be made of it, since the use may affect decisions as to how, when, and from whom the data should be collected. Most of our particular effort in AFHRL is directed toward development of some reasonable target against which we may validate our predictor tests. Historically, we have used technical school grades for this purpose, but the Air Force is rapidly moving to self-paced training, which poses very serious and rather obvious difficulties for psychologists who are charged with the development of selection procedures. Anyone concerned with the development of criterion instruments must be concerned with problems in the use dimension. We have all seen criterion ratings collected which were a hodge-podge of attempts to evaluate a person's current status, his future potential, and his past performance all rolled up willy-nilly into one exercise.

The use should be clarified and stipulated as early and as thoroughly as possible, and decisions taken at that point. For example, a criterion may be needed as a basis for rewarding past behavior. In that case, criterion information obviously should be limited to past behavior--ratings of potential are somewhat inappropriate. On the other hand, management may want to know which of several candidates is most likely to perform well in some new job which has opened up. In that case, ratings of potential would be preferred (incidentally, notice that ratings of potential are not really criteria; in the traditional sense--they are predictors of future performance, even though ratings are used to collect the information). Or perhaps the reason for collection of the data may be to decide whether or not to train particular employees. If so, perhaps a comparison (not a conglomeration) of current accomplishment and potential would be in order. The point is that a whole constellation of problems revolves around the uses to be made of criterion information, and that a great deal of thought should be given to the projected use of the information and the time line of intellectual development before the first step is taken to collect the data.

Level of complexity.

Still another dimension of measurement problems is created by the fact that intellectual development proceeds from more simple to more complex.

1. The economics of rating attractiveness. It takes longer and longer to observe all the necessary performance elements the further one moves along the continuum of intellectual development, since learning builds upon learning and current status consequently becomes more and more complex. This is perhaps the primary reason why ratings have been used and will continue for a long time to be used so prominently in the collection of criterion information.

If one is measuring complex behavior with tests, he must be prepared to require his subjects for longer and longer test sessions. One can measure physical strength, reaction time, visual acuity, and other simple characteristics in only 2 or 3 minutes each. It takes about a half-hour to get a reasonable measure of verbal ability. It would probably take at least 2 or 3 days of testing to get an adequate sample of behavior which would indicate a subject's proficiency in, say, aircraft engine repair. Indeed, we have seen reports describing some proficiency tests that require up to 11 days to administer (McKnight & Butler, 1964).

One assumes that a rater has already observed the complex behavior of interest for several days, and, given the proper conditions, he can report it with some objectivity. There is great appeal in an assessment metric which can be collected with no cost of subject time and very little of supervisor time. We have not yet been willing to pay the price of obtaining more objective and more accurate test data, so we sacrifice the greater objectivity of tests for the great convenience of ratings. Furthermore, ratings can be collected on any level of complexity desired, and I suspect that is why rating data collected in one situation frequently predict rating data collected in another situation, despite our certain knowledge that most sets of ratings contain many flagrant errors. The very fact that ratings can be made of very complex behaviors, compared with tests, means that we can reduce the distance between D_1 and D_2 in our formula and thus reduce the very important effects of potential and energy not well measurable at the present time.

113

We do not contend that this is as it should be, but it appears that this is the way it is and will continue to be; so we believe a strong attack on rating problems is of prime importance. Some of the rating problems that come immediately to mind are:

a. How important are the old reliable problems, such as halo, leniency, and the like?

b. What kinds of factors or characteristics make the best rating medium? In what formats should they be cast?

c. Just as there are apparently individual differences in rater accuracy, are there also reliable individual differences among ratees which affect the accuracy of ratings made on them?

d. Assuming that we can measure individual rater accuracy, what can be done in a situation using rated criterion data to improve the psychometric qualities of ratings collected from a mixture of both accurate and inaccurate raters?

e. We are convinced that if one intends to do research aimed at a better understanding of criterion variables, he must be prepared to do some social and organizational research as an integral part of his effort. Such a simple problem as a slippage in the worker-supervisor interface can cause very serious problems in performance evaluation. If the supervisor sees the job as primarily A, B, and C, and the worker sees it as primarily D, E, and F, the worker can be busy as an ant doing the wrong things.

We are studying all these rating problems, and we appear to be making a little progress.

2. Relevance. As one attempts to measure more and more complex behaviors, relevance becomes more and more important. Several investigators (Brogden & Taylor, 1950; Nagle, 1953) have pointed out the necessity of attempting to include all important elements of the criterion in the predictor set and to exclude from the predictor set all elements not present in the criterion. That, of course, involves a much more vigorous analysis of criterion variables than we are used to. But I am sure you are all familiar with relevancy problems, and they don't need to be restated here.

We see this set of problems as involving decisions about where and how completely to sample behavior along the line of development. For instance, it is likely that one who performs well on a test of mechanical aptitude will do well as an automobile mechanic if other conditions lead him to attempt the skill. A good automobile mechanic is likely to become a good carburetor specialist, and so on. If we want to find someone who will become a good carburetor specialist, do we measure his mechanical aptitude--which we can do quickly and easily--but which, by its nature, is too simple factorially to capture much of the variance we are interested in? Or do we measure his general automotive repair knowledge which is closer in time and in complexity to carburetor specializing but which is far more difficult to measure?

Questions of this sort have no easy answers. Trade-offs and compromise must be the order of the day until some breakthrough enables us to measure complex behaviors much more satisfactorily than we do now or until we learn how to use measures of simple behavior in a better, more comprehensive system.

One of the pitfalls we must be aware of is the seduction of a criterion just because it is there. Indeed, if the criterion metric is already there, just waiting for us to come use it, we should consider it immediately suspect. It is undoubtedly relevant for someone's purpose (or one assumes it wouldn't be collected), but it may have little or no relevance for whatever measurement concept the investigator has in mind.

To sum up, then, we believe that the little formula, $D_2 = D_1 (1 + i)^t$, and the line of intellectual development implied by the formula, has led us in some directions which we feel to be promising:

a. Because of the current difficulty of measuring complex behavior, we believe ratings will be relied upon for a long while to come. Because this appears true, we intend to concentrate a large portion of our resources on studying rating variance and trying to understand and correct for rating inaccuracies.

b. It would certainly help a great deal if we could plug in some solid values for the potential, the opportunity, and the energy which make up the term "i" in the equation, so that prediction of some point on the development line could be made with a more complete set of the simpler, more basic predictors. Some crude measures of all of these terms are already available, but a great deal of research needs doing, oriented around this point of view, to attempt to produce a more usable system.

3. A great deal of research needs doing on ways to measure complex behavior in an acceptable framework of subject time and overall expense. Some of our most strongly held psychometric ideas may have to be re-examined, particularly in our attempts to measure complex behavior. For instance, one cannot demand high internal consistency of items if he is attempting to construct a test which is deliberately complex. Indeed, it may well be that some techniques should be applied to item selection which simultaneously minimizes internal consistency and maximizes validity, such as the Horst Fan Technique or something similar.

Probably the formula is an oversimplification, but, whatever else the formula may have done or not done, we are certain of one value it has had for us. Though it may be illusory, it has at least contributed a little to our peace of mind as we grope our way through this maze of very complex problems.

REFERENCES

Brogden, H.E., & Taylor, E.K. The theory and classification of criterion bias. *Educational and Psychological Measurement*, 1950, *10*, 159-186.

Cattell, R.B. Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, 1941, *38*, 592.

Estes, W.K. Learning theory and intelligence. *American Psychologist*, October 1974, 740-749.

Ghiselli, E.E., & Haire, M. The validation of selection tests in the light of the dynamic character of criteria. *Personnel Psychology*, 1960, *13*, 225-231.

Horn, J.L. Organization of abilities and the development of intelligence. *Psychological Review*, 1968, *75*(3), 242-59.

McKnight, A.J., & Butler, P.J. Identification of electronic maintenance training requirements; development and evaluation of an experimental ordnance radar repair course (HumRRO RR-15). December 1964.

Nagle, F.G. Criterion development. *Personnel Psychology*, 1953, *6*, 271-289.

Otis, J.L. Whose criteria? Presidential address, Division 14, American Psychological Association, Cleveland, September 1953.

Prien, E.P. Dynamic character of criteria: Organization change. *Journal of Applied Psychology*, 1966, *50*(6), 501-504.

Thorndike, E.L. *Measurement of Intelligence*. New York: Teacher's College, Columbia University, 1926.

FOOTNOTE

Extraneous remarks by Dr. Mullins

1. To begin with, the paper that I'm going to give this morning is a purely speculative paper. This particular one simply describes our philosophy and ways that we have developed of looking at the criterion problem. There is nothing empirical in it; it's, as I say, just pure speculation. However, it does lead us to a point of view which has helped us quite a bit, and we hope it will help you.

XII

RATER ACCURACY

Joseph L. Weeks and Cecil J. Mullins
Personnel Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

Performance ratings have been in the past and probably will continue to be in the future the most common means of measuring job performance. The reasons for this are that they can be quickly obtained and are relatively inexpensive as compared to other techniques of measurement. Despite the frequency of their occurrence, there are many drawbacks to using performance ratings. Their typical low reliability and validity are generally recognized. Indeed, the measurement problems associated with ratings are so difficult that some researchers have suggested that they not be used at all (Ronan & Schwartz, 1971).

The basic problem with ratings lies in the fact that they actually represent second-hand accounts of performance. With paper-and-pencil devices, the subject records his performance on a piece of paper, a vehicle which is not subject to change, distortions, misunderstandings, poor memory, or gastrointestinal ailments. Such is not the case with ratings. The subject's performance is recorded in a particular situation, through a perceptual filter, on the memory of the rater and then, on some later date, is transferred to paper.

Apart from the difficulties associated with the performance evaluation process itself, rating research is often conflicting and repetitious. Evidently the reason for this lies in the fact that there is no generally accepted theoretical framework which serves as a guide to research. The majority of rating literature is devoted to the development of rating scales. Although the development of an objective, error-free rating scale is highly desirable, ratings are influenced by many variables, all of which deserve concerted research attention.

The rating paradigm, as we perceive it, consists of at least five basic dimensions: (1) At the top of the list is the rater. We know, for example, that his social adjustment, intelligence, similarity with the ratee, and position relative to the ratee will have substantial influence on ratings (Bruner & Tagiuri, 1954). There are probably many other rater characteristics associated with rater accuracy, as well. (2) The second dimension is the person rated. People differ in terms of the degree to which they can be accurately evaluated. Allport (1937) has indicated that some persons are more easily evaluated because they have more "open" personalities. Others, because they are more "enigmatic," are less easily evaluated. (3) To these dimensions can be added the traits or tasks to be rated. The value of judgments will vary depending on whether or not the traits employed have observable behavioral manifestations (Allport, 1937). Also, it has been found that the accuracy of ratings will decrease as the complexity of the task rated increases (Harris, 1966). (4) The social environment in which the ratings are collected will also have an effect. Kipnis (1960) indicates that leniency in ratings is more likely in a social environment described as supportive than one described as stressful. (5) Finally, the physical environment will influence ratings. Persons who are less observable due to arrangements of the work space will be more difficult to rate than those who perform in a situation that is more conducive to observation.

The last and perhaps most important consideration, although not strictly a rating dimension, is the purpose for which ratings are collected. The value of ratings will differ depending on whether they are collected for research purposes or for management decisions such as promotions and salary increases. The inflation of means and compression of variance typical of ratings collected for management decisions frequently eliminate them as useful criteria for purposes of test validation.

Obviously, the variables within each of these dimensions are quite complex. Considerations as to the manner in which interactions among these variables influence ratings boggle the mind. Our first research effort focuses on one of these dimensions, the rater. Specifically, it will be more concerned with the overall accuracy of judgments of behavior rather than with separate factors associated with rater inaccuracy such as central tendency, leniency, and halo. The goal of our research is to maximize the quality of rating data used for validation studies. If it were possible to identify the more accurate raters and use only their judgments, we would be in a considerably better position to determine the validity of our selection and classification instruments.

Scientific interest in the accurate rater or the good judge of personality occurred frequently in the 1930's and 1940's but eventually gave way to the investigation of rating errors. In an excellent review of the literature devoted to the ability to judge personality, Taft (1955) indicates that the ability to judge is related to intelligence, self-insight, emotional adjustment, and social skill. He further points out that accurate judgments are based on possessing appropriate judgmental norms, judging ability, and most importantly motivation. Recently, Gordon (1970, 1972) performed some very interesting research into the nature of rating accuracy. He suggests that rater inaccuracy is due to two types of errors, either "falsely accusing the ratee of doing something incorrectly which was in reality done correctly" or "giving the ratee credit for something that was actually done incorrectly." He provides evidence indicating a greater occurrence of the last type of error; that is, "giving the ratee credit for incorrect behavior," and concludes that the accuracy of ratings depends on whether or not the behavior observed is correct or incorrect.

The underlying assumptions for research into rating accuracy are that persons differ with respect to their ability to accurately assess performance and that there is consistency in their characteristic rating responses. Indirect research evidence is available to support these assumptions. Wiley (1959) and Wiley, Harber, and Giorgia (1959) reported studies based on rater's estimations of the qualifications necessary for various jobs. They concluded that rater differences do exist in a consistent enough fashion to justify their measurement.

A final rather critical assumption, which we will investigate, is that rater accuracy is a generalized ability. That is, we are assuming that the accuracy of ratings will be maintained across traits or tasks and ratees. Mullins and Force (1962) have gathered evidence which supports this assumption. Using a sample of inexperienced raters, they found that the capacity to evaluate verbal ability was directly related to the ability to evaluate carefulness. However, the statistical evidence obtained in support of this relationship was rather weak. In opposition to the assumption that rater accuracy is a generalized ability, Allport (1937) has indicated that "the ability to judge is neither entirely specific nor entirely general, but that it is probably more of an error to assume that it is entirely specific." Taft (1955) agrees and goes further to indicate that the validity of the assumption that rating accuracy is generalizable is dependent on a set of factors which include the subject rated, the traits employed, and the reliability of the criterion of accuracy. Since differences of opinion do exist as to whether or not this is a justifiable assumption, it is prudent to reserve judgment until further clarifying research has been accomplished.

Obviously, the major problem with research into the nature of rating accuracy is the establishment of a suitable criterion. That is, a more ultimate measure of the trait judged must be obtained and employed as a yardstick to determine the accuracy of the judgments made by various raters. In some research, pooled judgments of the rated trait have served as the basis for determining accuracy (Adams, 1927; Ferguson 1949; Greene, 1948; Wiley & Jenkins, 1964). However, as Taft (1955) has pointed out, with this technique there is the possibility that we are actually measuring the extent to which raters conform to the group consensus or display the same biases as the criterion judges rather than measuring rater accuracy. Other studies employed more objective criteria to evaluate accuracy. Vernon (1933) used a combination of independent ratings and test measures of the rated trait. Norman (1953) and Gordon (1970, 1972) measured accuracy in terms of the agreement between ratings and behavioral records. To circumvent the difficulties associated with using pooled judgments as a criterion of accuracy, we intend to use paper-and-pencil tests as a standard.

Our efforts will begin with a replication and extension of research performed by Mullins and Force (1962). In this study, differences between estimated and actual scores on a vocabulary test served as the criterion of rater accuracy. That is, subjects estimated their peers' scores on a vocabulary test after being informed of the average and range of scores for the group. For each rater, the differences between their estimates and the actual scores were averaged across ratees and served as the basis for classifying the rater as either accurate or inaccurate. It was hypothesized that if raters were correctly identified, the correlations between ratings of a behavioral trait (carefulness) and test measures of the trait would be greater for the accurate than for the inaccurate raters. The results of the data analysis supported this hypothesis.

In the extension of this study, we will manipulate the criteria used for identifying accurate raters. Differences between estimated and actual scores on a test of verbal ability and on a test of a less observable phenomenon, mathematics ability (and a combination of the two) will be investigated as a basis of determining rater accuracy. In addition, we will confirm our tentative identification of raters as either accurate or inaccurate on the basis of multiple traits. Not only will ratings and test measures of carefulness be compared as before, but also we will compare ratings and test measures of decisiveness, a trait less subject to observation than carefulness.

The last phase of the extension to the Mullins and Force study will involve an attempt to predict rater accuracy. Using averaged differences between estimated and actual scores on tests of verbal and quantitative ability as the criterion, we will determine the predictive efficiency of a set of variables hypothesized to be related to rater accuracy. The predictors will include measures of self confidence, gregariousness, surgency, and compulsivity.

The potential payoff for this type of research is great. Further down the road, we plan studies to determine if rater accuracy can be increased by training. In addition, we plan to investigate the possibility of statistically manipulating ratings in order to increase their accuracy. Obviously, we have just opened the lid on this type of research, and a lot of hard thinking must be accomplished to work out the details and overcome the obstacles. Nevertheless, we have confidence in this approach and feel that it will make a significant contribution to the state-of-the-art.

REFERENCES

Adams, H.F. The good judge of personality. Journal of Abnormal and Social Psychology, 1927, 22, 172-181.

Allport, G.W. Personality: A psychological interpretation. New York: Henry Holt, 1937.

Bruner, J.S., & Tagiuri, R. The perception of people. In G. Lindzey (Ed.), Handbook of Social Psychology, Cambridge, Mass.: Addison-Wesley, 1954, 2, 634-654.

Ferguson, L.W. The value of acquaintance ratings in criteria research. Personnel Psychology, 1949, 2, 93-102.

Gordon, M.E. The effect of the correctness of the behavior observed on the accuracy of ratings. Organizational Behavior and Human Performance, 1970, 5, 366-377.

Gordon, M.E. An examination of the relationship between the accuracy and favorability of ratings. Journal of Applied Psychology, 1972, 56(1), 49-53.

Guion, R.M. Personnel Testing. New York: McGraw-Hill, 1965.

Green, G.H. Insight and group adjustment. Journal of Abnormal and Social Psychology, 1948, 43, 49-61.

Harris, D.H. Effect of equipment complexity on inspection performance. Journal of Applied Psychology, 1966, 50, 236-237.

Kipnis, D. Some determinants of supervisory esteem. Personnel Psychology, 1960, 13, 377-391.

Mullins, C.J., & Force, R.C. Rater accuracy as a generalized ability. Journal of Applied Psychology, 1962, 46(3), 191-193.

Norman, R.D. The interrelationships among acceptance-rejection, self-other identity, insight into self, and realistic perception of others. Journal of Social Psychology, 1953, 37, 205-235.

Ronan, W.W., & Schwartz, A.P. Ratings as performance criteria. International Review of Applied Psychology, October 1974, 23(2), 71-82.

Taft, R. The ability to judge people. Psychological Bulletin, 1955, 52, 1-23.

Vernon, P.E. Some characteristics of the good judge of personality. *Journal of Social Psychology*, 1933, *4*, 42-57.

Wiley, L. *Determining job qualification requirements by rating Air Force task statements* (WADC-TN-59-41). Lackland Air Force Base, TX: Personnel Laboratory, Wright Air Development Center, July 1959.

Wiley, L., Harber, H.B., & Giorgia, M.J. *Rater tendencies in estimating qualifications required by Air Force tasks* (WADC-TN-59-195). Wright Air Development Center, September 1959.

Wiley, L., & Jenkins, W.S. Selecting competent raters. *Journal of Applied Psychology*, 1964, *48*(4), 215-217.

XIII

1. CONTENT ANALYSIS OF RATING CRITERIA

Eric D. Curton, Forrest R. Ratliff, and Cecil J. Mullins
Personnel Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

Introduction

For many years, much of the research concerning the content of evaluation instruments has focused on the relative merit of behaviorally-based and trait-oriented rating scales for the evaluation of job performance. One impetus for this research was the introduction by Smith and Kendall (1963) of a technique for the development of behaviorally anchored scales. Basically, the procedure entailed having people familiar with a particular job situation develop broad characteristics or factors which cover all aspects of the job. Behavioral examples are then developed to exemplify high and low performance points for each characteristic as well as moderate performance points within the two extremes. These behavioral examples are then written as expectations of specific behaviors and re-evaluated by independent judges. Only behavioral examples which are reliably judged as representing a particular level of performance on the same characteristic are included in the final evaluation form.

Since its introduction, the Smith and Kendall technique has been applied and evaluated in a number of settings both in the field and the laboratory. Its popularity is probably a result of the generally accepted viewpoint that it is psychometrically better to evaluate job performance using factors that are based on specific behaviors rather than factors based on personality traits.

The primary problem faced by someone trying to develop relevant performance factors for use in a large, complex organization is the time and expense involved in using something like the Smith and Kendall technique for the wide range of jobs encountered. The basic question that needs to be answered is whether objective, job specific factors are psychometrically superior to more subjective personal-trait factors in the evaluation of job performance. If the job-specific factors prove to be statistically superior, then the practical significance of the difference must be great enough to justify the cost involved in developing the more objective factors.

Relevant Research

In a review of the literature on the content of evaluation instruments, Kavanagh (1971) stated that the trend in this area of research has been toward the use of objective and measurable traits as opposed to personality traits in performance evaluation. He goes on to say that despite the fact that the objective traits were gaining in popularity, the empirical evidence in support of objective traits was not strong enough to warrant their use in exclusion of personality traits. Kavanagh further stated that the idea of an ultimate criterion of job performance is a behavioral construct and, therefore, construct validation should be the method by which immediate measures of performance are evaluated in terms of their relevance to the ultimate criterion. He then categorized the relevant literature according to the method of validation used in each study and reviewed them by category.

One group of studies used inter-rater or re-rating reliability as one method of validation. In general, the more objective traits proved to be rated somewhat more reliably, but the results were certainly not unequivocal, and many subjective personality traits also showed a high degree of reliability. Kavanagh points out that validity by consensual agreement is really a form of convergent validity and, according to Campbell (1960), both convergent and discriminant validity are needed for establishing construct validity.

Another group of studies reviewed by Kavanagh used validation against another criterion to determine the relevance of rating scale content. Kavanagh says that this approach is valid as long as the criterion used for validation is closer to the ultimate criterion than the ratings themselves. The problem is that this decision is usually judgmental rather than empirical. (This touches upon the problem mentioned in the paper by Dr. Mullins and Lt Col Ratliff with respect to differentiating between predictor and criterion and the fact that what we really have is a measurement problem.) In the group of studies reviewed, the more objective traits generally showed a somewhat higher validation against another criterion, but again the results were inconclusive. Some studies showed personal traits to be better than the more objective factors, and personal traits accounted for at least some of the variance in most of the studies.

The third group of studies reviewed by Kavanagh used validation by the multitrait-multimethod matrix introduced by Campbell and Fiske (1959). The use of this scheme allows one to obtain measures of both convergent and discriminant validity so that overall construct validity of rating scales can be better inferred. The results of the studies reviewed again proved to be equivocal with both objective and personal traits being psychometrically superior in different situations.

In concluding his article, Kavanagh points out that based upon the current literature, no absolute decision can be reached with respect to the superiority of one type of rating factor over the other in all situations. Kavanagh recognizes the basic problem of the relative efficiency of objective traits versus the amount of time spent in their development when he says, "objective job-oriented traits seem at present to have a slight edge, but the problem of situational specificity and additional time question the practical usefulness of this purist approach" (p. 663).

Since the Kavanagh article, very few studies have been done which specifically compare behaviorally-based and personality-oriented rating factors. Campbell, Dunnette, Arvery, and Hellervik (1973) evaluated behaviorally based factors which were developed for department store managers using a modified form of the Smith-Kendall technique. They found that when the factor scales were anchored with behavioral expectations, the ratings showed less halo, leniency, and method variance than when only broad definitions of the factors were used. While personality trait factors per se were not used in this study, it does show the decrease in the efficiency of behaviorally-based scales when they are not anchored with behavioral expectation statements. The authors also mention that "the managers who developed these scales invested a tremendous amount of effort in the process" (p. 22).

Neither of the two major studies which specifically compared behaviorally-based and personality-oriented factors found reason to overwhelmingly support either type of rating scale. Burnaska and Hollmann (1974) compared three rating scale formats using analysis of variance techniques. They compared Smith-Kendall type behaviorally anchored scales and scales with the same dimensions but without the behavioral anchors just as Campbell et al. (1973) had done. Additionally, Burnaska and Hollmann compared both of those formats with scales made from a priori determined factors and no behavioral anchors.

Unlike Campbell et al. (1973), Burnaska and Hollmann found that behavioral anchoring did not enhance the psychometric properties of the systematically developed scales. While they did find that the Smith-Kendall scales were somewhat less susceptible to leniency error and allowed greater differentiation between ratees, they concluded that "there is no evidence for the superiority of one format" (p. 311). They based this conclusion on the fact that all three formats contained composite halo and leniency error leading to small interratee discrimination. This fact led Burnaska and Hollman to question the ability of even systematically developed scales to diminish raters' tendency to rate according to an overall motivational component similar to Spearman's "g" factor.

Borman and Dunnette (1975) studied essentially the same variables that Burnaska and Hollmann had studied. The behavioral scales were developed to evaluate the performance of Naval officers, and the a priori trait-oriented factors were those already in use on the Naval Officer Fitness Report. They found that the behaviorally-based factors with anchored scales were psychometrically superior to the other two rating formats on measures of leniency, differentiation among ratees, halo, and interrater agreement. However, the magnitude of the differences was small, only sometimes reaching statistical significance. The authors state that probably less than 5% of the variance in the dependent variables can be accounted for by differences in the rating formats. Noting the amount of time and effort required in developing behaviorally-based factors, the authors question the usefulness of the Smith-Kendall procedure if the scales are only going to be used for performance ratings. They conclude that "at present little empirical evidence exists supporting the incremental validity of performance ratings made using behavioral scales" (p. 565).

The consensus of the literature to date is about the same as it was at the time of the Kavanagh (1971) review. Behaviorally-oriented, job specific rating factors are generally shown to be somewhat psychometrically superior to the more subjective personality trait factors. However, even when the sytematically developed scales are shown to be more efficient, the differences between rating formats are usually small. A real question still exists as to whether the superiority of the job specific factors, although statistically significant, is of enough practical significance to warrant the time and effort involved in their development.

Current Research

The Air Force Human Resources Laboratory has recently begun a series of studies at the Air Training Command Noncommissioned Officers (NCO) Academy. The purpose of these studies will be to analyze the content issue in an Air Force environment. Of particular importance will be determining the operational impact of various psychometric differences in sets of rating factors. Hopefully, methodologies developed and analyzed in this particular setting can later be used to develop criterion instruments for a wide range of Air Force jobs.

The NCO Academy at Lackland AFB provides in-residence professional military education for Air Force NCOs in the grades of E6 and E7. The NCO Academy classes last for about 6 weeks. Typically, there are 135 students per class, and they are divided into 9 seminars with 15 students in each seminar.

The general strategy of the studies will be to have the students at the NCO Academy render ratings on the other students in their seminar group. Means, standard deviations, pooled variance, and other traditional analyses will indicate the degree to which the rating factors are subject to rater errors such as leniency and halo. Also, the instructors will be asked to rate the students so that the convergent and discriminant validity of the factors can be determined by use of the multitrait-multirater matrix.

In addition to the traditional analyses done to determine the psychometric properties of the factors, profiles will be made up on each person based upon his or her average rating on each factor. These profiles will be returned to the students, and they will be asked to identify the people in their seminar groups from their profiles. They will also be asked to rank order the profiles according to how well they think a person with a particular profile will perform at the NCO Academy. Analysis of these data will show the number of times each person correctly identifies a classmate from his profile of scores. Also, correlations will be generated to show the degree of association between the rank ordering of the profiles and the actual rank ordering of students at the end of the class. These additional analyses will yield some measurement of the practical significance of differences in psychometric properties of rating factors.

Thus far, two studies have been completed at the NCO Academy. The first was a pilot study to determine and correct methodological problems that would be encountered. The most significant result from the first study was the identification of a set of 10 rating factors which the students agreed upon as being appropriate for evaluating their performance at the academy.

The second study has recently been completed, and the data are currently being analyzed. Table 1 shows the results of some preliminary analyses that were compiled from the data. While these results are in rough form and need to be analyzed much more thoroughly, they do give an example of the type of information that might be gained with our experimental design.

In this particular study, three sets of 10 rating factors are being compared. Two sets of factors come from a survey which was sent to Air Force NCOs in the grades of E7, E8, and E9. These NCOs were asked what factors they thought should be used to evaluate them on their jobs. The top 10 factors and the bottom 10 factors chosen by survey respondents make up two of the sets of factors used in this study. The third set of factors is made up of those factors chosen by the students at the academy as being appropriate for evaluating their performance. Each set of 10 rating factors was assigned to 3 of the 9 seminar groups at the academy. The students then used a rating form containing those 10 factors to rate the other members of their seminar group. They rated each student with each factor using a 5-point scale labeled "Far Below Average," "Below Average," "Average," "Above Average," and "Well Above Average."

Using mean ratings across all factors as a measure of leniency error, Table 1 shows that ratings using the student generated factors were less susceptible to leniency error than either of the survey generated factors. Of the survey generated factors, the bottom 10 factors were superior to the top 10 factors. This same relationship appears when considering the standard deviations of the factor scores, which is an indication of the degree to which the ratings differentiate among ratees. These are the types of analyses appearing in the literature today, and sometimes differences as small as those shown in Table 1 are used to support the superiority of one type of rating factor over another.

Table 1. Comparison of Three Sets of Rating Factors

	Student Generated	Survey Top Ten	Survey Bottom Ten
Means	3.56	3.74	3.63
Standard Deviations	.42	.33	.40
Hits	3.42	2.17	2.68
Correlations	.43	.42	.39

The next step in this study was to develop a profile on each person based upon his or her mean ratings on all factors. These profiles were then returned to the students, and each student was asked to identify the other students in the seminar group from their profile scores. In Table 1, "hits" are used to designate the mean number of times people were correctly identified using each of the three sets of factors. It can be seen that students using the student generated factors averaged identifying 3.42 out of 15 seminar members correctly while those using the survey bottom 10 factors identified 2.68, and those using the survey top 10 factors identified only 2.17 correctly. This analysis gives an indication that the relationships shown with the mean and standard deviation scores have an influence on how well people can be separated and identified in an operational sense.

If differentiation among ratees were the goal of the rating instrument, then it appears that the student generated factors are superior to the survey bottom 10 factors which are in turn superior to the survey top 10 factors. It also appears that the measurement of means and/or standard deviations of the factor scores would give a reliable indication of the relative superiority of the sets of factors without going through the identification step.

However, simple identification and differentiation is rarely the goal of a rating instrument. Instead, it is usually used to judge how well a person performs his job. If a rating instrument did give an accurate assessment of how well a job was performed, then differentiation among ratees would certainly be achieved, assuming the ratees performed the job at different levels of ability. However, even though differentiation among ratees should result from using a valid rating instrument, the fact that differentiation occurs is not sufficient evidence for the instrument to be considered valid for evaluating job performance. A good example is shown in the present study.

The students were asked to rank order the profiles according to how well they felt a person with a particular profile would perform while at the NCO Academy. Table 1 shows the average correlations between the rank ordering of the profiles and the actual rank ordering of the students at the end of the class based upon their final grades. It can be seen that the differences between correlations are insignificant and that one set of factors seems to be just about as good as another for actually predicting the performance of a ratee. Therefore, while one set of factors is psychometrically superior to another set, when judged against the criterion of actual job performance, the superiority of any one set of factors disappears. This seems to point out the importance of these additional analyses in trying to determine the relative effectiveness of a set of factors in an operational setting. While one factor or one set of factors may be psychometrically superior to another, the practical significance of the differences should be investigated before an operational decision is made.

REFERENCES

Borman, W.C., & Dunnette, M.D. Behavior-based versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology, 1975, 60, 561-565.

Burnaska, R.F., & Hollmann, T.D. An empirical comparison of the relative effects of rater response biases on three rating scale formats. Journal of Applied Psychology, 1974, 59, 307-312.

Campbell, D.T. Recommendations for APA test standards regarding construct, trait, or discriminant validity. American Psychologist, 1960, 15, 546-553.

Campbell, J.P., Dunnette, M.D., Arvey, R.D., & Hellervik, L.W. The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 1973, 57, 15-22.

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validation by the multi-trait multi-method matrix. Psychological Bulletin, 1959, 56, 81-105.

Kavanagh, M.J. The content issue in performance appraisal: A review. Personnel Psychology, 1971, 24, 653-668.

Smith, P.C., & Kendall, L.M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

2. THE NORMATIVE USE OF IPSATIVE RATINGS

Cecil J. Mullins and Joseph L. Weeks
Personnel Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

Whenever ratings are collected from supervisors in an operational setting, particularly if the ratee must be made aware of the rating given to him, two undesirable consequences usually occur. The ratings become "inflated" (that is, the mean approaches the upper range limit), and the variance becomes compressed (that is, everybody gets essentially the same score). The major reason these two effects occur is that the supervisor is required to perform mutually incompatible acts--he must support his people and he must critically evaluate his people. It is very difficult to do both, so the reaction of most supervisors, at least in large organizations, is to try to see that his people get a better than average chance at promotion. As a consequence, ratings creep up and accuracy falls off.

The effects just mentioned occur when operational ratings are collected normatively. Normative scores are those which produce norms, so that comparisons may be made across individuals in a group. A ratee's score may be expressed as a percentile, showing his standing in relation to other members of the group.

There is another kind of data which can be collected in a manner that automatically minimizes the inflation of means and the variance compression customarily found when normative data are used for operational ratings. Rating data can be collected in a manner (called "ipsative" ratings) such that characteristics within an individual are rated relative only to other characteristics of the same individual. This method produces a profile of the characteristics, showing which of the ratee's traits are his stronger ones and which are his weaker ones. Nothing can be inferred about the strength of any of the ratee's characteristics, as compared with the strength of some other ratee on that characteristic. If a list of characteristics is ranked for a particular ratee from strongest to weakest, there is absolutely no problem with mean inflation and variance compression because the mean and the variance are fixed mechanically by the ranking process.

However, ipsative rankings (relative rankings of characteristics within the individual ratee) are useless for operational evaluative purposes unless they can be treated in some way so that the information on each ratee can be compared with that for other ratees. For example, it does little good to know that, say, creativity is Joe's strongest characteristic and Mary's weakest characteristic if we are trying to compare Mary with Joe. It is entirely possible that Joe is generally so inept and Mary so generally expert that Mary's creativity, although it is her weakest characteristic, may still be stronger than Joe's creativity, although it is his strongest characteristic.

We can see two ways to convert ipsative rating data so that comparisons can be made across individuals. One of these ways is by computing an index of worker-job match. It is obtained simply enough by correlating the ranking of characteristics describing the individual with a similar ranking of the same characteristics as they are required by the job, as shown in Figure 1. The ranking of job characteristics should be performed by someone other than the one who ranks these characteristics in the worker. The correlation coefficient may be used in raw or converted form as an index of worker-job match. It seems likely that if two workers are of the same level of general competence averaged across separate applicable skills and traits, the one whose pattern of characteristics most closely resembles the pattern required by the job will be the one who performs better. The worker-job match index can be included with whatever other variables are available as candidates for criterion composites.

	Rankings	
	Mary	Job X
Carefulness	1	3
Responsiveness	2	1
Initiative	3	4
Creativity	4	5
Tolerance of stress	5	2
Cooperation	6	9
Adaptability	7	7
Writing ability	8	10
Speaking ability	9	8
Reasoning ability	10	6

Rho = .72

Figure 1. The computation of a worker-job match index.

The worker-job match index yields information which should prove useful. However, another treatment is possible, and we plan to investigate it. A worker's pattern of characteristics could correlate perfectly with the pattern required by a job, but he could be so generally weak that he performs poorly; or he may possess such all-around competence that he does well despite a poor job-match index.

All the job-match index reveals is the congruence of patterns of characteristics between the worker and the job. It provides no information at all on the relative strengths of two workers on any of the characteristics. This is not a serious problem if the worker-job match index can be included as just one component in a composite criterion along with at least one pertinent normative variable. The normative variable will establish a level of general competence, and the worker-job match index will be weighted to the extent that pattern congruity is important. But there are some situations in which tests are disliked as a means of worker appraisal. In these situations, if only one test can be administered or if a score from a previously administered test can be obtained from the files or if any kind of reasonable normative variable is available on a large number of workers, then a situation can be set up so that an anchoring system can be employed. The anchoring variable is common to the workers being evaluated and is ranked along with the other characteristics. The other ranked characteristics will fall above or below the anchor variable according to how they are ranked for a particular worker. Standard scores (percentiles, z-scores, or something similar) can then be assigned to each of the ranked characteristics so that comparisons can be made across individuals on each of the characteristics.

The conversion to standard scores required for this approach was mentioned glibly in the previous paragraph, as if the problems surrounding this important step were all solved. They have not been. We believe we can produce a crude system of conversion now, but it will need much sharpening. The production of standard scores such as these involves some knowledge about intra-individual variability across characteristics. We know that there is a fairly strong tendency for positively regarded characteristics, both intellectual and non--intellectual, to be intercorrelated, (Horn, 1968). To the extent that these characteristics are correlated, to that extent the intra-individual variability will be reduced, and the more accurately standard scores can be assigned to the ipsatively ranked characteristics. Our first cut will be a very primitive conversion system based on distributions of intra-individual variability obtained on other groups and other characteristics (see Figure 2). The standard scores issuing from this conversion system certainly will not be exact, but they should be accurate enough to yield evaluations which, because of their relative immunity to deliberate biasing by the supervisor, should prove more useful than the system ordinarily used.

These standard scores will then be in a normative form, and they become possible candidates, appropriately weighted, to form a composite criterion score. The weights would be obtained by using the variables as predictors of some more ultimate criterion, or of some criterion which may be obtained experimentally but not operationally. It should be obvious that the anchor variable system is not substantially different from a system using the worker-job match index in conjunction with at least one normative score on an appropriate variable. We plan to compare both these systems.

Characteristic	Pete's Ranking	Percentile
Carefulness	1	85.5 [75 + (7/10 x 15)]
Responsiveness	2	
Initiative	3	Range (from studying other characteristics, other populations) = 15 percentile points
Creativity	4	
Tolerance of stress	5	
Cooperation	6	
Adaptability	7	
Reasoning ability	8	75 (measured anchor variable)
Writing ability	9	
Speaking ability	10	72 [75 - (2/10 x 15)]

Reasoning ability test score = 75th percentile.

Figure 2. Calculation of normative values for ipsative rankings, using an anchor variable.

Perhaps you will remember from the line of intellectual development we discussed yesterday that it is our conviction that there is no single criterion, immutable and all-encompassing. There are innumerable points of intellectual development from birth to death, each a little more complex than the previous one. It is conceivable that each of these points may be eventually measurable, but each is so complex that it is unlikely that any point ever will be completely measured for any practical purpose other than research. A criterion is a measure, taken at a desired point along the development line, of that portion of intellectual development which seems to the investigator to represent those functions with which he is most directly concerned. That point may serve both as a criterion for predictors consisting of earlier points and as a predictor for criteria taken at later points. With this orientation, it is quite reasonable to "validate" criterion measures against other criterion measures.

Because of the nature of this system, many studies will have to be done before we can say with any confidence that the system is worth the effort. The following questions, and many others, will have to be answered:

1. Is the proposed system a better way of collecting evaluation information than the simpler one of collecting normative rating data? It appears that it should be better, but one cannot know for sure until the system has been subjected to empirical scrutiny.

2. The efficiency of any evaluative scheme depends in large part on the particular variables selected to enter the system. What is the

best way to select the variables needed? Captain Curton addressed this problem in his presentation.

3. What weights should the various components of the system take? For example, is the worker-job match index the most important consideration, or the least important, or somewhere in between?

These short statements of research questions actually involve very long and very difficult research work. We don't know how good the system will prove to be, but we believe that it should at least be better than the system of collecting rating data which is currently used so widely.

REFERENCES

Horn, J.L. Organization of abilities and the development of intelligence. Psychological Review, 1968, 75(3), 242-259.

XIV

SYNTHETIC CRITERIA

Cecil J. Mullins, Forrest R. Ratliff, and James A. Earles
Personnel Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

Now and then a predictor battery is required in a situation where no criterion exists. This kind of situation can arise when a new specialty is born and there are no subjects currently performing in the specialty; or when the specialty is so thinly manned or unusual that requisite numbers of performers for validation studies simply do not exist; or when management needs a predictor battery substantially sooner than one can be produced by the classical validation technique. Seven years ago, AFHRL developed two methods for furnishing a using agency with a predictor battery immediately upon request, if the using agency could provide a team of subject matter experts for about a half-day's effort (Mullins & Usdin, 1970). As part of the research work connected with this effort, a comparison was made between the battery furnished in the classical way and the batteries furnished with these two synthetic methods, and it appeared that there was no practical difference among the batteries in their efficiency in predicting an empirical criterion. The two techniques are called the R-technique and the M-technique, and both are based on the assumption that synthetic criterion vectors can be devised which are similar enough to the empirical criterion vector so that weights produced for the predictor variables in the synthetic criterion situation will be essentially the same as predictor weights generated in the classical empirical situation. The focus of our previous research was almost entirely on the utility of predictor weights produced synthetically, but we believe now that a good estimate of the empirical validity coefficient can also be produced synthetically. Both synthetic techniques make a few other important assumptions:

1. It is assumed that decisions have already been made, or can be made, about which predictor variables will enter the predictor battery. This means that the variables are available off the shelf, or that the preliminary work on the variables (concerning item analyses, reliabilities, etc.) has already been accomplished. The predictors are ready to go--all that remains is the problem of relative weights for the separate predictor variables.

2. It is assumed that the requesting agency can furnish at least three subject matter specialists who are thoroughly conversant with the demands of the job to be performed, and that the producing agency can

141

furnish at least three test specialists who thoroughly understand the tests in the predictor battery, or who can be made to understand them by a brief statistical description of their characteristics.

3. If one is doing research on the techniques, it is assumed that some empirical criterion will be available so that the weighted composite scores generated synthetically can be compared for efficiency with the weighted composite score produced empirically. If one is not doing research, but simply producing a battery for a using agency, this assumption is not absolutely necessary, but empirical demonstration of the degree of efficiency of the synthetic composites is still desirable if a criterion can be obtained. In the latter case, obviously, the synthetically produced prediction composites can be considered as a stop-gap measure until empirical weighting becomes a possibility.

R-Technique

The R-technique requires that the subject matter specialists and/or the test experts (the judges) rate 100 subjects on how well the judges believe, from studying the subjects' scores on the predictor variables, the subjects will perform on the job of interest. The 100 subjects need not be real people--they can be made up. If they are real people, they should be selected from available subjects in such a way that considerable spread is introduced into the profiles which are studied by the judges. When the 100 subjects have been rated, the ratings are used as a criterion against which all the predictors for these 100 subjects are correlated. The multiple correlation, of course, produces a set of weights for the predictor variables which are then used to calculate a predictor composite for each of the subjects one is really interested in.

M-Technique

The M-technique is also a way of arriving at relative weights for the various predictors, so that a prediction composite can be calculated for the subjects of interest. The judges also provide the information for this technique, but the information is of rather a different kind. Instead of estimates of likely performance of a sample of dummy subjects, the M-technique produces estimates of relative importance of variables comprising the predictor set. The predictor variables are factor analyzed, the resultant factors are explained to the judges, and the judges are told to distribute 100 points among the factors according to how important the judges believe the factors are in producing good job performance.

If a real criterion were available, it could be introduced into the factor analysis and its correlations with any predictor would be reproducible by multiplying the criterion's factor loadings by the corresponding factor loadings of the predictor and then summing these products across all factors. In this way, a validity vector can be produced from a table of factor loadings. But our problem involves a situation where no criterion exists.

Since no criterion exists, and consequently no criterion factor loadings exist, the square roots of the distributions of 100 points among the factors by the judges must substitute for the loadings. Then, by the arithmetic described above, an estimated validity vector is produced and, from this, weights for the various predictors are obtained. The details of both techniques for producing weights are contained in the Mullins and Usdin report.

In the previous work done on these techniques, a criterion of technical school grades was available for 1,000 subjects from each of four schools, one in each of the Air Force's four aptitude areas (mechanical, administrative, general, and electronic). An empirical composite was computed in the usual way. Each of the four samples was randomly split into two 500-man subsamples. One of these subsamples was used to generate weights, and the other was used to cross-validate. The cross-validated R was used as a reference point, and, within each of the four cross-validation subsamples, other prediction composites were computed for each subject, generated by the synthetic approaches. In most instances, the synthetically generated composites produced validities which, for practical purposes, were not different from those produced in the usual empirical way. In only one school was the prediction of the empirical criterion significantly worse using the synthetically generated composites, and that difference was barely significant at the .01 level.

At the present time, two further investigations of these techniques are under way. One of these investigations is analogous to the previous study in that technical grades are once again the criterion of the prediction battery. The other on-going investigation expands the application of the techniques to the prediction of ratings of on-the-job performance.

If the replication work currently under way produces results as encouraging as the previous study, this approach to validation of our Air Force predictor tests will form at least an interim position while the search for a satisfactory criterion continues.

REFERENCES

Mullins, C.J., & Usdin, E. Estimation of validity in the absence of a criterion (AFHRL-TR-70-36). Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, October 1970.

XV

WHAT IS THE VALUE OF APTITUDE TESTS?

Raymond E. Christal
Occupation and Manpower Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

Introduction

The title of my paper is "What is the Value of Aptitude Tests?" No one could feel comfortable dealing with such a broad and controversial topic--especially in front of a group of professionals in the testing business--but I feel the topic needs to be discussed and debated.

Recently, some individuals have gone so far as to suggest that testing be done away with altogether. Good heavens! Haven't we demonstrated for decades the value of tests in personnel selection and classification? Of course we must deal with reasonable questions concerning the fairness and job relevance of tests, but surely all military managers should see that tests are indispensable.

Evidently, we have done an inadequate job in merchandising our product. For this reason, I would like to look at the manner in which we have attempted to see the value of tests and see if there are holes in our case. Then, I will venture to make a few suggestions for reorientation of our sales pitch and research strategies.

Present Defense

As I review the situation, I find that we have defended the value of aptitude tests on three grounds: (1) their ability to predict performance on the job; (2) their ability to predict attrition in training; and (3) their ability to predict course grades. I would like to consider these one at a time.

Prediction of Job Performance

First, let's consider job performance. Now let's be honest about it. We really don't have overpowering evidence that our tests predict job performance, and informed managers and operators know that we don't. Many of these individuals are of the opinion that the key to productivity is not individual differences in aptitude, but good management.

Experience teaches them that nearly all personnel they deal with on a day-by-day basis could get the job done if they simply applied themselves. The individual differences they observe are mostly motivational, or else are not job related.

Of course, these managers are right. What they fail to understand is that this lack of variance is, to a large extent, the product of testing and training. If managers in an electronics maintenance occupation were to receive a random sample of untrained personnel out of the general population and attempt to generate the required skills on the job, I can assure you that they would quickly become acutely aware of individual differences in aptitude. However, this would not be an efficient way to run a military service. We use tests to select and classify individuals into occupations such that each person has the capacity to acquire the necessary skills for acceptable job performance. The training program, in turn, is geared to provide each trainee with these required skills. If the process is efficient, then there is no reason why tests should predict performance variance on the job, and we should neither make apologies nor hang our heads in shame when such is found to be the case.

Prediction of Attrition

The second way we have defended our tests is by showing how well they predict attrition in training. In the Air Force, a washout in pilot training costs the service thousands of dollars, and the claim is made that millions of dollars of additional costs are avoided each year by using tests to screen out applicants likely to fail in training. On the surface, this sounds like a strong case for tests. It can be shown that within any training class, individuals with high aptitude scores wash out at a much lower rate than individuals with low scores. It is also true that washouts are very expensive. However, it is not easy to demonstrate that our aptitude tests save money by reducing washout rates.

Let me show you some data extracted from the *Army Air Forces Aviation Psychology Research Report No. 2* (DuBois, 1947).

Table 1. Attrition Rates and Aptitude Input for Every Third Pilot Training Class (44C thru 45G)*

Class	N	Aptitude Cutoff	Percent Eliminees
44C	12,232	3	15.5
44F	9,371	3	12.0
44I	6,466	4	19.6
45A	6,525	4	21.0
45D	1,384	4	21.5
45G	664	6	27.4

*Extracted from Report No. 2, "The Classification Program," *Army Air Forces Aviation Psychology Program Research Reports*, 1947.

Table 1 reflects pass/fail data for every third class from 44C through 45G. In classes 44C and 44F, the cutting score on the aptitude score for entry was Stanine 3, and the average attrition rate was 13.9%. In classes 44I, 45A, and 45D, the cutoff was raised to Stanine 4. However, instead of going down, the attrition rate increased to 20.4%. Finally, in class 45G, the cutting point was raised to Stanine 6, yet the attrition rate went up again--clear up to 27.4%. In view of these data, one might conclude that attrition in pilot training would be minimized if those cases having the least aptitude were entered into training.

Of course, this is not true. The fact is that attrition rates were controlled by administrative actions, and were not dependent on the quality of the input. The number of pilot graduates was determined in large part by the number of cockpits to be filled. The data shown in Table 1 reflect actions taken toward the end of the war as the number of trained pilots became abundant and aircraft production was reduced. We have good reason for believing that the quality of graduates from these classes varied, but we cannot demonstrate that the use of tests saved money by reducing attrition rates.

We would have even a more difficult time demonstrating the influence of tests on attrition rates in enlisted courses. The number of graduates from such courses is ordinarily programmed months in advance to meet operational requirements, and fluctuations in input talent produce only minor fluctuations in attrition rates. During periods of low quality input, it is not uncommon to increase wash-backs and remedial training to maintain production standards.

Pass/fail is a very slippery criterion, and attrition rates seem to be arbitrarily established. This phenomenon is not restricted to the military. For example, there are wide variations in the input talent to colleges and universities, where attrition rates for the same courses are essentially equivalent. A washout from MIT or Cal Tech could be an honor graduate from certain other colleges and universities. We seem to be living in a relative world without absolute standards. This is one of the problems we face in demonstrating the value of tests.

In 1957, Dr. Krumboltz and I published a study (Krumboltz & Christal, 1957) in which we demonstrated that the probability of a student completing pilot training is a function of the aptitude levels of the other three students with whom he is grouped under the same instructor. A student with a Stanine 5 was less likely to graduate if he were grouped with three students at the Stanine 9 level than if he were grouped with three students at the Stanine 5 level.

In 1959, an investigator in Australia reported a strange and related finding (Want, 1959). In that country, Air Force and Navy pilots were being trained together under the same instructors. The Air Force raised their entrance requirements, and the result was that the attrition rate for Navy trainees nearly doubled. While the level of talent of Navy trainees remained constant, these individuals began looking bad in comparison with their Air Force counterparts.

These studies demonstrate that aptitude tests do measure differences in abilities which are recognized by instructors. However, we will not be able to defend our tests on the basis of their role in reducing attrition rates until absolute standards for successful course completion are implemented and adhered to.

Prediction of Course Grades

A third way we have attempted to show the value of tests is in terms of their ability to predict final course grades. The statement that aptitude tests predict course grades is irrefutable. Literally hundreds of studies have consistently demonstrated this to be so. To prove that we haven't lost our grip in this respect, I've brought along results from one of the largest Air Force validation studies ever conducted, which I will display to you.

We began with a 380,000-case population graduating from Air Force entry-level courses between January 1969 and April 1974. From this population, we randomly selected 1,000 cases from each course, when available, or a total sample when data were available from fewer than 1,000 cases. This yielded a total validation sample of slightly more than 100,000 cases, representing graduates from 134 different courses.

Table 2. Validities (R) of AQE/ASVAB/AFQT for Course Grades*
for AFSs with AQE/ASVAB Cutoff at 80th Centile

R	N	R	N
.626	59	.422	1000
.543	303	.422	1000
.507	679	.421	1000
.485	749	.414	1000
.483	168	.409	1000
.483	426	.407	988
.472	1000	.406	1000
.471	1000	.396	152
.471	249	.395	503
.471	1000	.394	217
.463	1000	.387	1000
.457	509	.386	1000
.456	1000	.383	1000
.444	1000	.382	753
.439	624	.379	637
.438	1000	.374	1000
.437	1000	.366	209
.435	1000	.348	716
.431	1000	.324	1000
.429	608	.285	283
.427	777	.164	1000
		Median R = .424	Total N = 31,609

*For cases graduating between Jan 1969 and Apr 1974.

The validity coefficients I will show are uncorrected multiple correlation coefficients for a weighted composite of the four AQE composites and AFQT against final course grades. The values in Table 2 show the validities computed in 42 courses for which the cutting score on AQE was at the 80th centile. These coefficients may look a little low, but remember that they are uncorrected and have been computed in a sample which has been subjected to severe restriction in range on the predictors. Since the bivariate normality assumptions could not be met, no corrections for restriction were made. However, it is estimated that in an unrestricted population, many of these validities would be found to be in the .60s, .70s, and .80s. The median correlation obtained in the computing sample was .42. The lowest reported validity is for a Linguistic/Interrogator course for which the Air Force has special additional screening procedures.

Table 3. Validities (R) of AQE/ASVAB/AFQT for Course Grades*
for AFSs with AQE/ASVAB Cutoff of 60th or 70th Centile

R	N	R	N	
.647	78	.439	210	
.631	139	.439	1000	
.624	658	.435	1000	
.619	163	.422	697	
.586	434	.415	823	
.551	1000	.410	146	
.535	605	.405	129	
.531	1000	.392	412	
.529	606	.389	1000	
.527	908	.386	999	
.527	332	.385	1000	
.518	1000	.381	425	
.518	1000	.370	114	
.517	1000	.348	1000	
.502	892	.327	1000	
.498	612	.305	1000	
.492	1000	.305	1000	Comp Operator
.491	65	.232	228	Comp Programmer
.484	539	.176	202	Small Arms
.474	1000	.173	1000	AC&W Operator
.458	291	.158	1000	Radio Operator
.440	1000			
		Median R = .440	Total N = 28,707	

*For cases graduating between Jan 1969 and Apr 1974.

Table 3 reports uncorrected validities for 36 courses having entry-level requirements at the 60th or 70th centile on AQE. Again, these coefficients are attenuated by severe restrictions in range, although some of the uncorrected Rs are higher than .50.

I might point out that five of the lowest six coefficients in this table are associated with courses training students in operator-type jobs. Two are for radio and morse system operators, for which a special code test is available to enhance prediction of student success. The other three are for computer operators, aircraft control and warning operators, and small arms specialists. In each instance, certain perceptual-psychomotor skills are required which are not measured by the AQE or AFQT.

The median uncorrected validity of the tests for these 42 schools was .44 which, again, is a gross underestimate of values which would have been obtained in an unrestricted sample.

Table 4. Validities (R) of AQE/ASVAB/AFQT for Course Grades* for AFSs with AQE/ASVAB Cutoff of 40th or 50th Centile

R	N	R	N	R	N
.678	807	.557	532	.488	1000
.672	636	.556	1000	.482	628
.668	105	.552	437	.479	1000
.657	100	.550	177	.465	850
.652	140	.549	641	.465	1000
.634	980	.544	1000	.440	1000
.628	1000	.542	1000	.432	305
.625	649	.536	666	.432	814
.592	1000	.535	240	.422	1000
.591	1000	.432	1000	.412	1000
.588	532	.531	1000	.404	890
.584	886	.528	598	.392	1000
.581	1000	.527	376	.378	609
.574	1000	.521	1000	.375	1000
.572	1000	.498	208	.371	1000
.570	715	.493	563	.369	191
.568	1000	.490	1000	.351	372
.566	1000	.489	751	.263	1000
.565	575	.489	1000	.221	1000

Median R = .532 Total N = 42,973

*For cases graduating between Jan 1969 and Apr 1974.

Table 4 reports validities for grades in 56 courses for which AQE entrance requirements are at the 40th or 50th centile levels. These coefficients are higher because they are less subject to restriction in range. The median value is .53. However, these coefficients are considerably below what would be obtained in an unrestricted sample. Not only have the lower 40 to 50 percent of the standardization population been denied entry into the course, but the number of cases in the upper levels of the aptitude distribution is severely limited due to siphoning off by more demanding courses.

Once again, by the data I have presented, we can demonstrate that aptitude scores predict course grades. I'm not sure, however, that this fact impresses the average military manager. After all, one cannot translate course grade points into dollars and cents or manpower bodies; nor have we been able to demonstrate convincingly that graduates with high course grades actually perform better on the job than graduates with low course grades, even though they in fact may do so.

Summary of Current Status

So here we stand. Although we feel that aptitude tests predict job performance, we have very little data to support this contention. We would like to claim that the use of tests reduces attrition in training, but the evidence suggests that attrition rates are primarily a function of administrative actions, not level of input talent. We can show that test scores predict course grades, but this doesn't seem to impress the average military manager. Where do we go from here?

Suggested Criteria for Test Evaluation

It would be my recommendation that, in the future, we focus our attention on five types of criteria for test evaluation as follows:

1. Speed of skill acquisition
2. Speed of skill decay
3. Speed of skill reacquisition — Skills Maintenance
4. Speed of response
5. Accuracy of response — Performance

Speed and accuracy of response may be important in some occupations involving a demand for perceptual-psychomotor or clerical skills. However, due to time limitations, I have elected to address only the first three criteria, which relate to the *speed* of skill acquisition, decay, and reacquisition. In all three instances, the basic variable against which tests are to be evaluated is *TIME*. Time is an excellent criterion. It has a zero point; it can be measured in equal intervals; it is easily understood by military managers; it can be easily converted into dollars and cents or manpower spaces; and it is the single most expensive item in the military budget.

The military services spend literally billions of dollars each year supporting the development and maintenance of skills. The more obvious expenditures are associated with formal residence and on-the-job training courses, but this is just the top of the iceberg. For example, the Air Force spends hundreds of millions of dollars each year just to maintain pilot and navigator skills. Even more costly is the time individuals in all services spend in learning to perform new tasks as they are encountered on a day-by-day and assignment-by-assignment basis. To the extent that aptitude scores predict the time required for individuals to acquire and maintain skills, they can be used to reduce costs and optimally distribute talent to jobs. I will address this issue during my remaining time.

Skills Acquisition

There is nothing unique or new about the concept of aptitude scores predicting learning rate. For example, in 1963, John B. Carroll recommended that aptitude be defined as learning rate (Carroll, 1963). The first intelligence test developed by Alfred Binet, back in 1904, was designed to measure differences in the level of skills acquired by individuals during a constant time interval (chronological age). These scores were later normed and converted into a score "mental age." A ratio of the mental age to chronological age was computed and came to be called the Intelligence Quotient (IQ). Regardless of the problems associated with the development and utilization of IQ scores, they have been used for years as rough indicators of individual learning rates.

In the academic world, many tests are called learning abilities measures, and have been used for decades by teachers to place pupils into homogeneous groups so as to minimize variance in learning rates within groups. Tests have been shown to be valid predictors of school grades, both in the academic world of the civilian sector and in all military services, and school grades can be viewed as the amount of content mastered by students when learning time is held constant. Aptitude tests also predict proficiency test scores in the services, which are rough measures of the amount of content mastered by individuals at various career points. In Project UTILITY (Vineberg & Taylor, 1972), which was conducted for the U.S. Army by the Human Resources Research Organization in the late 1960's, AFQT scores were shown to be related to the rate of skill acquisition in several occupational areas. However, with the passage of time, an increasing proportion of men at all levels of AFQT appeared in the upper ranges of performance distributions, indicating that for these low-level occupations aptitude scores predict the rate of skills acquisition, but not ultimate level of performance. Pilot training programs are generally locked-step. For this reason, I have been unable to locate data demonstrating that aptitude scores predict speed of skill acquisition. However, pilot aptitude tests do predict within-class elimination for flying deficiency, and individuals in the flying research area assure me that slowness in acquiring skills is the primary cause for such elimination. This observation needs to be confirmed by carefully controlled research.

While the evidence that aptitude scores predict learning time is substantial, most of it is indirect. Outside of a few laboratory experiments dealing with paired associates learning, I have been able to locate few studies directly addressing the subject, and these have involved small N's and produced mixed results. In one study conducted by a graduate student at the University of Pittsburgh (Wang, 1968) and in another study conducted by the Human Resources Research Organization (Wagner, Behringer, & Pattie, 1973), substantial relationships were found between general and specialized aptitude tests and learning times; however, there appeared to be complex interactions among learning rates, types of materials to be learned, training modalities, and various aptitude scores. If such findings are generally confirmed, the proper selection and classification of personnel may be more complicated than it appears on the surface. However, in one unpublished study conducted by the Navy,* no such interactions were found, and standard Navy aptitude tests were demonstrated to have substantial validity for predicting training times (see Table 5). This study involved two tracts in a Navy aviation familiarization course, one which was made up solely of reading modules, and the second which included seven slide/tape modules. Interestingly, the higher validities were obtained for the slide/tape group. Notice that the equations predicting time criteria for the two treatments were highly homogeneous.

I was also able to obtain data for a 200-case sample of Air Force personnel who recently completed an individualized instruction course (Inventory Management) at Lowry Air Force Base. Two criteria were available, one of which was a summation of time to complete the course blocks, and the other of which was a summation of course block scores (grades). The results of this analysis are presented in Table 6. The multiple validity of the ASVAB composites and AFQT for the training time criterion was only .39--which was significant, but lower than hoped for. However, the multiple validity of three ASVAB composites for the sum of block test grades was .59, which is higher than was obtained for final school grades when the course was taught in a locked-step fashion. Even though this course is now taught in an individualized instruction mode, there appears to be more predictable variance in the amount of content mastered than in the time for course completion. This finding is explained, in part, by the fact that students in the course took module and block tests when they felt they were ready for examination. Upon first testing, some students barely

---

*Information in this table was provided by Dr. Kirk A. Johnson, Navy Personnel Research and Development Center, Memphis Branch Office, Millington, Tennessee. Multiple R's and cross-application R's were computed by the author using the correlation matrices provided by Dr. Johnson.

152

Table 5. Validities of Aptidude Scores for Time (Hrs) Criteria in Navy Aviation Familiarization Course

Group #1 - 7 Slide/Tape + 9 Reading Modules (N = 109)

Aptitude Test	Validity
GCT	-.58
Arithmetic	-.47
Clerical	-.34
Multiple R	-.67

Group #2 - 16 Reading Modules (N = 113)

Aptitude Test	Validity
GCT	-.45
Arithmetic	-.43
Clerical	-.26
Multiple R	-.51

Multiple R's and Cross-Application R's

	Development Sample R	Cross-Application Sample R
#1	.67	.66
#2	.54	.53

Table 6. Validities of ASVAB/AFQT Scores for Time and Grade Criteria in the Air Force Inventory Management Course (N = 200)

Criterion	Development Sample R's Predictors	Multiple R
Time	General AI, Electronic AI, AFQT	.39
Grade	General AI, Electronic AI, Mechanical AI	.59

Cross-Application R's

Source of Predictive Weights	Application Criterion	R
Time Criterion Predictors	Grade	.37
Grade Criterion Predictors	Time	.55

reached a 70% passing standard, while others routinely scored 100% on many tests. These latter students had reached the 70% standard at much earlier (but unknown) points in time, so there was no simple way to compute a time-to-standard for each case. In this sample, the correlation between the time and grade criteria was -.40, indicating that students completing the course in the shortest time tended to be those who mastered the greatest amount of content.

There is not time to discuss problems associated with generating a pure time-to-standard criterion in the operational setting, but I would like to recognize that such problems do exist. It is unlikely that individualized instruction courses presently train all students to exactly the same standard (although some meet a 90-90 standard), even though finishing times may vary. Until this problem is resolved, it will be difficult to establish the exact relationship between aptitude scores and learning rates in such courses. Ultimate solutions may include better records and controls, continuous testing, statistical corrections, and controlled experiments. One must admit that the problems to be overcome are challenging.

It should be observed from Table 6 that the equations predicting the grade and time criteria are homogeneous. This provides additional evidence that, since tests normally have high validity for course grades, they should also be found to be highly related to learning time criteria. It is important, however, that direct relationships be established. The author would appreciate receiving copies of any studies bearing on the question.

Prediction of Decay Rates

A second stream of research which needs to be initiated concerns the ability of aptitude tests to predict decay rates for skills and knowledges. There has been a great deal of research leading to the development of generalized curves of retention, but surprisingly little research has been accomplished relating to individual differences in retention. Underwood published one summary paper (Underwood, 1954) in which he concludes that, when associative strength is held constant, there are no differences in forgetting rates as a function of aptitude during the first 24 hours. However, this study dealt with laboratory associative learning experiments and short decay periods. The military services should be able to provide more definitive answers concerning individual differences in forgetting rate as a function of aptitude.

One very revealing study was reported by the Naval Personnel and Training Research Laboratory in 1970 (Johnson, 1970) which provided data relating to the skill decay question. The study was based on material being taught in the first phase of the avionics fundamentals course. Proficiency was measured by means of the criterion referenced tests that had been used to validate the programmed instructional material used in this phase. Measures were obtained on a pre-test, on an immediate post-test, and at intervals of 1 day, 7 days, 28 days, and 96 days following the original learning. It was found that in spite of a fairly high level of mastery on the immediate post-tests and a considerable amount of review, much of the material learning during the first phase of the course was forgotten by the end of the course. The differences between individual students were large on the pre-test, were quite small on the immediate post-test, and increased gradually over the remaining post-tests until, by the end of the course, they were almost as large as they were on the pre-test.

Although this study was based on only a fairly small N, it did provide a set of relatively unique data. The experiment began with 141 students. Seven were dropped for administrative reasons; 8 failed because of slow progress; 21 washed back because of slow progress; and 17 were moved ahead because of fast progress. Thus, only 85 cases were left in the final sample, and these cases were fairly homogeneous in terms of learning rate. In spite of this homogenization process, data in the study can be re-analyzed to reflect differential decay rates as a function of aptitude. As can be seen in Table 7, aptitude scores account for 24% of the final test score variance, with *original pool test scores held constant* (partial multiple R^2). Although one might argue that associative strength was not held constant, from a practical standpoint it can be stated that individuals showed differential decay rates in criterion referenced test scores as a function of their aptitude levels.

Table 7. Retention of Electronics Fundamentals as a Function of Aptitude

	Validities for Final Post-Test	
Predictors	R^2	R
Immediate Post-Test	.185	.430
Aptitude Tests Alone	.312	.559
Immediate Post-Test Plus Aptitude	.382	.618
Unique Contribution of Aptitude Tests	.197	.444
Aptitude Tests with Immediate Post-Test Scores Held Constant (Partial R^2 and R)	.242	.492

Predicting Time for Reacquisition of Skills

The third area which needs to be addressed concerns the time required for reacquisition of skills and knowledges which have degenerated over time as a function of disuse. One would hypothesize that if aptitude scores predict the speed of skills acquisition, they should also predict the speed of skills reacquisition; but, to my knowledge, this has not been firmly established in the military setting. I conducted one analysis in the early 1950's which I now wish I had documented, since it bears on the question. A number of World War II pilots were recalled during the Korean conflict and sent to flight instructors' school. At the school, they were given training to re-establish their flying skills. I managed to locate the original World War II pilot aptitude scores for a sample of these individuals and found, to my amazement, that they were still predictive of flying proficiency grades for students in this course--in spite of the passage of time and in spite of the original screening, training, and differential experiences these individuals had during and subsequent to World War II.

The question concerning the relationship between aptitude and the time required for skills maintenance is extremely important. For example, consider the pilot area alone, where the Air Force spends hundreds of millions of dollars per annum in terms of fuel, aircraft, and maintenance costs in order to maintain flying proficiency. In the foreseeable future, multi-millions of dollars will be spent for sophisticated simulators in hopes of saving fuel and aircraft associated with this expensive but necessary program. Yet, we know very little about the rates of skill decay and regeneration, and practically nothing concerning individual differences in such rates. Are individuals who quickly attain pilot skills also those who slowly lose such skills and quickly regain them after decay? If so, proper selection of individuals into the pilot training program may be more important than generally recognized. Because of the large numbers involved, the potential savings might be even larger on the enlisted side, although they may be more difficult to document.

Summary

I realize that I have wandered far and wide in this rather loosely organized paper, but I will try to summarize briefly. I have suggested that we should begin moving away from job performance, pass/fail criteria, and school grade criteria for aptitude test evaluation. Certain types of perceptual-psychomotor tests and tests of clerical speed and accuracy may predict performance in operator and clerical type jobs; however, we should not expect tests to have predictive efficiency for performance in jobs where performance is primarily a function of the extent to which fully developed skills are applied. Test scores do predict the relative probability of failure within training groups, but they do not determine failure rates for groups as a whole. Pass/fail rates are determined by administrative actions, rather than quality of input. Test scores predict training grades, but grade points cannot be easily translated into dollars and manpower.

I have suggested that we should demonstrate the value of tests in terms of their ability to predict personnel time requirements for skills acquisition and maintenance.

Finally, I have enumerated some of the research findings to date which bear upon critical issues, and have suggested research studies which should be undertaken.

I am personally convinced that aptitude tests are indispensable in the military setting and that they must continue to be utilized in spite of problems which may exist with respect to test fairness. I have faith that ways will be found to eliminate or reduce test biases which may exist. At the same time, I feel that we have an obligation to demonstrate the value of tests in terms of their ability to help us operate our military establishment in a cost-effective manner.

What is the value of aptitude tests? I cannot give a precise answer to this question; but they are of considerably more value than most military managers have been led to believe.

REFERENCES

Carroll, J.B. A model of school learning. Teachers College Record, 1963, 64(8), 723-733.

DuBois, P.H. (Ed.). Army Air Forces Aviation Psychology Program Research Report No 2. Washington, DC: U.S. Government Printing Office, 1947.

Johnson, K.A. Retention of electronic fundamentals: Differences among students (Technical Bulletin STB 71-2). Naval Personnel and Training Research Laboratory, October 1970.

Krumboltz, J.D., & Christal, R.E. Relative pilot aptitude and success in primary pilot training. Journal of Applied Psychology, 1957, 41, 409-413.

Underwood, B.J. Speed of learning and amount retained: A consideration of methodology. Psychology Bulletin, 1954, 51, 276-282.

Vineberg, R., & Taylor, E.N. Performance in four Army jobs by men at different aptitude (AFQT) levels: 3. The relationship of AFQT and job experience to job performance (HumRRO Tech. Rep. 72-22). Alexandria VA: Human Resources Research Organization, August 1972.

Wagner, H., Behringer, R.D., & Pattie, C.L. Individualized course completion time predictions: Development of instruments and techniques (HumRRO Tech. Rep. 73-25). Alexandria VA: Human Resources Research Organization, November 1973.

157

Wang, M.L.C.C. *An investigation of selected predictors for measuring and predicting rate of learning in classrooms operating under a program of individual instruction*. Doctoral dissertation, International University Microfilm, 1968.

Want, R. The frames of reference of flying instructors. *Journal of Applied Psychology*, 1959, *43*, 86-88.

SUMMARY AND CONCLUSIONS

Editor's note: The panel of invited experts were asked to comment on the specific papers, presented here under "Consultant Comments," and to provide closing summaries, included under "Summary Statements." Additionally, since these comments were off-hand and verbal, each consultant was later invited to prepare and submit a more formal paper giving his impressions of the symposium. Those papers received in response to this invitation are published together in the last section, entitled "Impressions."

159

CONSULTANT COMMENTS

Dr. R. Campbell: I was interested in the discussion of the combined ipsative and normative approach to rating and I was curious as to the projected purpose.

Dr. Mullins: Well, the primary purpose is to reduce the inflation of means and to increase the variance. You have to get it. Whether this variance is meaningful variance we won't know until we try.

Dr. Brokaw: The problem is that we're trying to determine whether the selection and classification variables we've been using are appropriate for that task.

Dr. R. Campbell: Okay, you can see other uses for such a measure, but if it's restricted to that I guess it helps clarify it for me. But I think the work of Mike Beer at Corning Glass was interesting in this regard. Are you familiar with what he's done?

Dr. Mullins: No.

Dr. R. Campbell: It's not published yet.

Dr. Mullins: Maybe that's why I'm not familiar with it.

Dr. R. Campbell: He's spoken about it someplace where I happened to be and it will be published soon (Personnel Psychology). He started out with an ipsative approach and his purpose was multi-faceted, it was not only focused on validation--I'm not even sure he had that in mind--but ran into the same problem. He needed an anchor because management rejected the ipsative approach. It didn't tell them enough for administrative matters. His anchor turned out to be an overall rating of performance. The whole anchoring issue raises real questions about the utility of the ipsative approach and whether or not it's really going to yield anything. I find the most attractive aspect of the ipsative approach to be for feedback to individuals on a diagnostic basis about their performance. Beyond that, I have difficulty seeing how it will be very helpful, particularly when you seem to be moving in the direction of g, away from a number of dimensions.

Maj Sellman: I have just a straightforward descriptive question on the number of people who have talked about doing work on job performance measurement via simulation as a real training, that sort of thing. I was wondering if you could, from the various branches, give some estimate of how many lives that's really touched, that is how many people to whom it has been applied, and just how widespread is it.

160

Col Ratliff: Mr. Camm has gone.

Dr. Muckler: I could give you some fourth-hand information from a paper I heard at AERA in April on that, and they were talking about how they implemented it. If I can remember right, I think they had a sample of 150 in each of two divisions over in Germany, and it was on an experimental basis but from what I heard in New York that was the extent of it at that point--the tryout over there. They'd sent a rather large number of researchers over to Germany to do it. I don't know how wide it's gone beyond that, but I know they are going to follow it up quite a bit.

Maj. Sellman: Is there anybody in the Navy who has to go through simulation training?

A: Where simulation is used as a measure of performance, I'm sure 50,000 people a year in the Navy are subjected to this.

Q: How many different jobs does that encompass?

A: 50,000.

Q: Is that done during training, post-training, or both?

A: Both. Post-training use of simulation and associated job performance measurement within the Navy is increasing constantly. If you ask me how well we're doing it, I would prefer not to answer that.

Dr. Muckler: If you don't mind, I'd like to stick a summary comment in at this point and come back later. I've been somewhat bothered by the frequent reference to the expense and the impracticality of work samples, simulations, and the like. I would like to point out to somebody in the Air Force (and I have a feeling that the people I would like to point this out to are not here), that the price of one B-1 bomber would be more than adequate to do an enormous amount of work on the development of practical, useful work samples. I would also like to point out, and this time, I think, to the people that are here, that there has been one area of confusion in the discussions here. That is that there has been almost intermingled discussion of performance measurement as research criteria and performance measurement for operational purposes. If you're concerned about a criterion measure, you're concerned about research work, and I do not believe that it is necessary or even desirable to use operational measures of performance as research criteria. The practicality of the work sample approach to performance measurement ought not get confused between the practicality of its use as a research tool and the practicality of its widespread use throughout the service as an operational tool. I was kind of startled--I'm going to even quote the sentence in Mr. Foley's paper when he made the statement--"At the present time throughout his whole career, a maintenance specialist is not required to demonstrate on formal job task performance tests that he can perform efficiently and effectively the tasks of his job." I think this is shameful, because I'm reading more into it than was actually said and I think I'm justified in doing it if the Air Force is anything like any other organization I've ever worked in. And if there are no formal job performance tests that an individual ever has to demonstrate proficiency on throughout his entire career, there probably is no systematic means of evaluating that performance either. We live in a society that worships hardware, that puts all of its faith in hardware, and that pays very little attention to the cost of the human organism that built the hardware, maintains the hardware, and operates the hardware. And until we get the notion that it is not practical to build all that hardware without giving some attention to the people that use it and do something with it, we really aren't going to be talking about anything very practical. End of sermon.

Dr. Brokaw: He's not here to defend himself, so I can pick on Ray Christal a little bit. If I can read my notes I can pick on him. He identifies speed as the all purpose criterion and level as the all purpose predictor. Now that suggests that a lot of people are wasting a lot of effort in a lot of places. I would like some individual and consensus responses to this concept. Do you think that this could be an artifact because he worked on groups which are already separated in terms of classification? He looked at mechanical people in the context of other mechanical people, he looked at electronics people in the context of other electronics people. He has not yet looked at these people in competition with each other...... Did I put everybody to sleep?

Dr. Hutchinson: I'd be glad to respond but not to that question.

Dr. Guion: It seems to me that--this is going to be on the tape so Ray can hear it, isn't it? Okay Ray, here we go. It seems to me that what he's done--what you have done, Ray--is to move back to World War I when we got all those beautiful charts that were reproduced in every elementary psychology textbook for a period of a generation or more showing the mean and standard deviation of AGCT scores for various occupational groups. I've always found that diagram to be one of the more interesting and useless diagrams in elementary psychology textbooks. Students spend a great deal of time pouring over it trying to decide which occupation has the intellectual prestige to which they aspire, but I have never found any practical usefulness for it in a non-military setting. If you go as far as Ray went and identify the crucial problem for military services being a placement or classification problem rather than a selection problem, I think that the oversimplicity of this model becomes so obvious that it no longer has any interest. Wish you were here, Ray.

Dr. J. Campbell: I would add a brief comment to that. I think Bob was saying appropriate things about the aptitude distributions for different occupations. However, I think that is a separate issue from whether the time it takes to reach an acceptable level of job proficiency is a useful criterion for selection and classification research.

Dr. Guion: I'm only talking about the general level as the generalized predictor.

Dr. Helmick: I would like to use this opportunity to raise a general question and apply it to this particular situation. It seems to me that one of the things that I saw getting lost in the discussion over the 2 days was the distinction that Dr. Muckler tried to make between measurement and criterion and the concept of the judgmental aspect that goes into what I would agree is the real, true aspect of the criterion. It seemed to me that the speed determination, as I understood it to be described, was essentially another measurement and really had nothing to do with the definition of the criterion. And I think it's a quite appropriate way under certain circumstances to measure the criterion. It may very well in many cases be a better way. Where you have mastery criteria, speed may very well be the only alternative. But that doesn't answer the basic question of speed to do what. How did you decide to measure the speed to acquire this particular kind of performance? It seems to me that a great deal of the discussion this morning as well as yesterday was concerned with measurement problems. I'm certainly not averse to that. Measurement problems are very real. But I think sometimes we stay in our difficulties because while we do refine the measurements, we still may not be measuring what we would like to if we stopped to think about it.

Dr. McCormick: Perhaps in defense of Ray Christal in his absence here, I would like to say that I believe that his position regarding "level" requirements for jobs does have a fair amount of validity to it. In other words, I think there is some tendency for people to gravitate into the kinds of jobs which are commensurate with their own levels of ability. Those persons who have that which it takes to perform a particular job may well perform at a different level on some test or other measurement instrument than persons on other jobs. I think in some of our research we have some evidence to support this. The assumption that people generally gravitate into jobs that are commensurate with their own levels of abilities is not a completely valid one, but at the same time I think that there is enough substance to this notion to support Ray's point that "level" of performance on various kinds of tests may be a reasonable criterion for the selection or placement of people on the jobs in question. With respect to the matter of "time" to learn various jobs that he discussed, I think basically the notion of time does make a certain amount of sense, although it does not completely avoid the business of making some kind of determination about the level of proficiency. In other words, to determine that the time required to achieve a certain level of proficiency one still has to make a determination as to the level of proficiency that you are talking about, so you do not completely avoid the business of evaluation, rating, or performance appraisal, or what not by the use of time. In connection with this matter of time, Stanley Lippert (whom some of you people may know) recently turned out a very thorough analysis of learning curves in which he has found some generalizable curves in which he has incorporated provision for measurement of the level at which a person begins learning whatever it is to be learned. On the basis of his evidence, I think that if time is used as a criterion, there should be some provision for incorporating a measure, at the initiation, of the performance level at which the person begins the training in question.

Dr. J. Campbell: I don't know if I can add anything to what's been said, but it seems quite reasonable to expect that as the military services move toward more self paced training, some good criterion measures to consider would be the time to training completion and the time to reach job proficiency. Another useful criterion might be the amount of decay in job skills after a certain amount of time. However, none of these gets one out of the bind of having to measure performance itself. Without measures of job performance, and a good definition of what constitutes an adequate performance level, it would not be possible to determine the time it takes an individual to reach "adequate performance." Thus the development of a criterion based on time will be more, not less, complicated than the usual kind of performance assessment. However, I'm sure this is not news to Dr. Christal and that he well realizes the difficulties involved. I think his argument is that, in spite of the difficulties, time is a very valuable criterion for military organizations. I also think he is right. However, perhaps without meaning to, he rather quickly slid over the problems that will be involved in rating the time demands for various job tasks. It won't be easy and it adds another rating task to whatever is already required of whatever sample of raters is available.

On the question of how to classify or place individuals in different Air Force jobs, I don't think I was able to fully understand what was said and thus should not comment on it. Nevertheless, I think some of us inferred that he was advocating a return to job placement via differential score levels on one dimension of overall ability. However, I don't think he would take such an

164

extreme position. We didn't hear correctly.

Another aspect of the general problem that seems missing from the discussion so far is that some job tasks are more "critical" than others, and predicting the time to learn the critical tasks would be more important than predicting the time to learn the less critical tasks. Another feature of the criticalness of tasks, which was recognized in a study of Navy enlisted personnel by Glickman and Vallance, is that there are often a finite number of identifiable ways that people fail at a job. That is, it is often possible to describe, in concrete behavioral terms, the most important mistakes that people make. If the objective is to select people who will minimize such mistakes then perhaps the most appropriate criterion is not the time it takes to perform the tasks adequately but the absolute level of proficiency with which an individual can learn to perform the task, given a reasonable amount of time.

Dr. Brokaw: We've had a lot of discussions of ratings. We've talked about ipsative ratings, we've talked about normative ratings, and we've talked about doing away with ratings in favor of performance tasks, and yet we seem almost always to come back to look at them again. I would like for you gentlemen to tell us whether we should go our merry way with ad hoc ratings as they seem to be appropriate or should we spend some time on attempting to develop some specialized rating kind of processes whereby we either train raters to levels of proficiency, or we identify raters who have success in the skill of rating objectively, or, what should we do about this rating problem. Should we assume that all the problems are answered, or should we pursue our research in that domain?

Dr. R. Campbell: I can give you a brief answer to that as I think ratings will be with us throughout my lifetime; however, I was encouraged by the emphasis on proficiency measurement (as distinguished from performance measurement) and I applaud that work. If you've got proficiency measures for 50,000 jobs, I think that's marvelous. We substitute them for proficiency ratings whenever possible in my organization. The fact is though, we will need ratings for other purposes. Now I certainly hope we would not use ad hoc ratings. Somebody here said we shouldn't use them, I think maybe several people did. I'm not very big on "rater accuracy" as the way to go. Frankly, it's an unfruitful way. I prefer improving rating conditions, and the training of raters, particularly if we're using these ratings in research situations. I think much can be done to make the ratings better.

Dr. Helmick: I would certainly agree. I think that from all of that I was encouraged by the attention that's being given to improving ratings, although I do not disagree that any time we can find a better measurement than a rating we ought to use it. I guess the only specific point I would raise in connection with the report on some of the work being done would be the emphasis, as I understood it, on trying to validate ratings against paper-and-pencil tests. Coming from one of the largest suppliers of paper-and-pencil tests in the world I certainly have no objections to them, but I have the feeling that modifying the rating procedure to produce results more like the paper-and-pencil tests would not necessarily be an advancement in approaching the truth. The kinds of things that can be effectively measured by paper-and-pencil tests may be less useful than those for which ratings may be the only means available.

Dr. McCormick: I think there are two kinds of circumstances under which ratings will continue to be used. In the first place, there are certain kinds of job activities which by their nature I believe can best be evaluated on the basis of subjective judgments of other people. As an example, in the case of behaviors of interpersonal nature, human judgments about such activities might be better than any other kind of measure. In the second place, ratings will, of course, have to continue to be used in the case of things that theoretically at least can be measured objectively but that we have not been bright enough to figure out how to measure. Now, as we think about what we call "ratings," I prefer really to think of the kinds of responses which "raters" are required to make. In the case of conventional ratings, the rater is asked to make absolute judgments, as contrasted with the making of relative judgments, when we talk about what I sometimes call personnel comparison systems (like rank order, forced distribution, paired comparison, etc.). I am in accord with Ray in his talk about the use of relative ratings. I think the notion of ipsative ratings also falls into this ballpark too. I think that the use of relative ratings can get around some of the problems of inflation and bunching up. However, there are other kinds of "rating procedures" that do not require the making of judgments or evaluations, but rather that require descriptions of behavior. I am thinking here of various types of scales and checklists such as behavioral expectation scales, the forced choice checklist, etc, where the "rater" is asked more to describe someone's behavior, rather than to judge or evaluate. I would heartily endorse any efforts to make comparisons of the effectiveness of these different kinds of human responses, both in terms of their psychometric properties and also in terms of their practical utility in connection with the whole matter of criterion development.

Dr. J. Campbell: In general, I guess one could say that any research on ratings is valuable. However, there are certain kinds of research that make me more nervous than others. It seems to me that research efforts devoted to discerning the value of different scoring procedures, different formats, different transformations, etc., is not really the direction to take. The historical record here is clear. These kinds of variables don't seem to make very much difference in the reliability and predictability of ratings. If you want to choose the one thing that makes the most difference, I think it is the motivational contingencies under which the rater operates. Also, I don't think the suggestion to be more descriptive than evaluative will help much. People (raters) know the purpose for which the ratings are being made and they know the rewards and punishments that are contingent on their behavior as raters. These motivational concerns are a significant influence on how they use the rating instrument. If we don't deal with these concerns then we don't deal with one of the major problems involved in the evaluation of one person by another. In my opinion, a second major determinant of reliability and accuracy in ratings is how well the raters understand the content of the behavior to be rated. Back a decade or so ago when Smith and Kendall stimulated our interest in the method of Behavior Expectation Scaling, people showed an interest in this technique for one of two major reasons. Some saw it as a way for sampling and describing job behaviors in a more complete and meaningful way than has ever been done before. For others it was a new way of dealing with the traditional problems of unreliability, halo error, leniency, etc. I think research on the BES method got on the wrong track early by emphasizing the latter and not the former objective. People should worry more about the "goodness" of the description of job behaviors to be rated and not so much about halo or leniency. In sum, I want to assert that two major areas of needed research are the motivational considerations influencing rater behavior (for research as well as operational ratings) and ways in which domains of critical job behaviors can be better and more usefully described for the raters.

Dr. Brokaw: Could the members of the panel comment on types of rater training programs they might have encountered like with the police department? Do you ever come across programs where they literally train you or somehow try to get the rater to make more accurate, valid, reliable, or useful ratings, more meaningful ratings?

Dr. Guion: Let me stick that into the more general comment that I want to make. I was going to let Mac speak for me on the rating issue until John started to confuse the operational ratings with the research ratings. And the thing that I think has to be recognized with regard to the research ratings is that even when you take the punishments and the rewards out of the thing and you tell the raters that what they're really doing is making it possible for them to get better people in the future or something of this sort and that nobody's going to get hurt or helped by their ratings in this particular set of ratings, they still can't do it. And this is true when we've given video tapes and training programs to them in a wide variety of different kinds of efforts. We have used films of actual police calls, for example, and gone through a great deal of intensive effort to get people to observe, describe, evaluate, agree on the meanings of anchor terms, this sort of thing; we still end up with many raters giving us terribly unreliable ratings even in that wholly laboratory situation where there isn't even the reward system of the research ratings.

I think that one of the things you have to recognize in responding to the question that was originally raised, is not merely that ratings will always be with us, but that they are ubiquitous. I think that we would do better if we stopped using the term "rating" and used the more general term instead, of judgment. We would recognize then that all the rating systems that we use as criterion measures, whether they are ratings per se or ratings of product or process in a work sample, or the evaluations that are made when someone is given a trial period of performance on a job, such as a probationary period, whatever the context in which the criterion measure exists, the rating is simply a tool for obtaining judgments.

The paper by Uhlaner, Drucker, and Camm has one interesting statement in it that would be interesting to question them about to see if they really mean it quite as it sounds. It offers the hypothesis that ratings are more likely to be "accurate" in those situations where some kind of inter-personal activity is involved. That's an interesting hypothesis. If this is true, then we should be using not only the whole process of judgment and perception research in our research on ratings, but we should be specializing perhaps on social judgment theory with all of the lens model implications, policy capturing implications, that this sort of thing has.

I guess the answer that I will have to give to your question, Lee, is that on the way down here I was reading these papers in the same week in which I had the first draft of two theses, one of which I've already told you about; it was the prediction of rating accuracy study that I mentioned. The other one was an interview study where we tried to determine the effect of non-verbal cues on interviewers' judgments which was a rather devastating kind of non-finding when we got all through with it. This, coupled with the fact that I happen to be at a university that has been specializing in the person of one of its faculty members in social judgment theory for the last (how many years, Jack, 8? 10? something like that!), I have made a vow to have a mid-career change and devote most of my attention over the next few years to the whole process of judgment in the evaluation of anything, performance, product, consequences of behavior, whatever you like--because I think that most of our criterion measures ultimately involve the process of judgment. Certainly they involve the process of judgment if we make the distinction that was urged upon us yesterday between a measuring instrument and the value judgment that turns the measuring instrument into a criterion measure. How these judgments are arrived at, the multiplicity of policies in arriving at those judgments, all of these are of crucial importance if we're going to evaluate criterion measures. And I don't think that we can simply walk away from ratings even as we walk toward job samples, simulations, and that kind of thing.

Dr. J. Campbell: Although this notion is not original with me, I think there is a law of nature that says objective measures are really subjective measures, at least one step removed. Behind every objective measure one can turn up personal judgment somewhere, and all the problems inherent in making such judgements come home to rest. That is why we all should be very concerned with problems of perception. Person perception research in social psychology, for example, has built up a huge literature on a lot of trivial things but also a lot of things which are very relevant for this situation. To mention just one, there is a large literature concerned with the influence of stereotypes on judgments. I can recall a study in the organizational literature by Wayne Kirschner which discovered fairly clearly, I think, that if you took two kinds of supervisors, those who were judged to be good supervisors and those who are judged to be bad supervisors, they had a very different stereotype of what a good employee was. As a result, one might expect them to rate different people highly or the same people differently. The person-perception literature is a big area and to be a well integrated investigator of problems of personal judgment (e.g., performance ratings), you must jump into it at some time or other.

Dr. Brokaw: Does anyone in the audience have a question?

Sgt. Winn: I've got a question. I'd like a quick summary of what the two different kinds of supervisors thought was a good employee.

Dr. J. Campbell: Well, a "quick" summary is that for the good supervisors, their stereotype said that a good employee was a little mavericky, a hard driver, a bit of a non-comformist, etc., whereas the poor supervisor's stereotype said that a good employee was docile, don't make too many waves, etc. I'm overstating the case a bit, but the descriptions were of that nature.

Dr. Guion: Our studies are of trained versus untrained leaders, but I think in most of them the training has to do with familiarizing people with what is halo, what is leniency, and where are all these so called psychometric errors. And it's very short. I have a paper here by one of Bowling Green's ex-students which compared trained versus untrained raters, where they were training, and this sort of thing. But I think the training that's really crucial is training in what are you going to observe, what are you going to call high, what are you going to call low, etc., and it's taking a long time. I mean we really train them.

Maj Sellman: In Flying Training, we have kind of a unique problem and that is to observe 10 minutes of behavior costs several thousands of dollars if you fly an airplane, so you want to make sure that whatever rater you have is the best possible rater, and pilots tend to be very good raters to start with. They really know what they're doing. But it's just a very difficult situation, and this is in a research area mostly. I'm not sure exactly what happens out in the--I mean just out there doing it.

Dr. Brokaw: There was a question Bob raised a while ago of how to get two people to agree that they've seen a specific behavior. It's no small problem.

Capt Curton: I just wanted to ask Dr. Campbell from AT&T if he would comment on the types of instruments they use to validate their assessment centers and promotions that result from the assessment centers; what types of criterion do you use in that situation?

Dr. R. Campbell: We have used several different types. One is advancement in the organization. The assessment centers are designed usually to show potential for advancement or potential for certain lines of work, and the criterion we have used most frequently is actual advancement in cases where the assessment center data was not fed back to the organization. So that's the most common one. Another is to set up special judgment situations where--I can think of a sales example where we were trying to validate an assessment program for salesmen--where there is a prescribed procedure for opening a sale, how you close a sale, how you do usage prospecting--getting information and so on, and there is a trained set of raters who normally go around the country doing evaluations of people so in that case we used a research judgmental procedure. Another approach that's been used with some success, at least in terms of showing validity, is instead of using the ratings that are in the files, administrative ratings, we have trained interviewers go out and talk to the supervisors who report on the behavior of the incumbent, and then the interviewer makes the judgment about where somebody falls on a certain dimension. Those are the main three--we've used salary progression but we avoid using ratings that are "available."

Col Ratliff: Dr. Campbell, in your assessment center where you use different people as part of your assessment process, do you think that participating in the assessment process makes a difference on their later ability to assess people?

Dr. R. Campbell: I believe so but I have no evidence to substantiate that. In staffing an assessment center, we do not rely on selection of good raters. We do provide training, explain the dimensions to be observed, and train them in observation and judgment. Perhaps the best training that they get in the whole thing is that they participate as a team of raters in which they have constant feedback on the judgments and observations that they're making. The problem when they get back to the field is while they may be better trained judges, I believe they certainly are, now the rating conditions are different. Now they don't have the full observation, they're not rating people on standard tasks, so whether or not their judgments are in fact better after they get in the field I really can't say although we think they're better trained raters.

Col Ratliff: You mean they're more conceited about their judgments when they get back.

Dr. R. Campbell: No. It's one thing to be trained in what kind of behavior you should be observing, what it applies to, and what the anchors are, but you've got to somehow set up the rating conditions so that you see the behaviors when you get out there. And you don't have that same control on the everyday field study that you have in an assessment center where everybody has gone through the same tasks. How much that impacts the "validity" of the judgments that they're making in the field I'm not sure.

Maj Waters: I'd like to just sort of comment. One of our divisions that's not represented here, our Flying Training Division, is doing a concerted effort in the performance evaluation area in the flying game, and they're specifically looking at automated performance measurement in the aircraft and pilot tasks. Since Jack was pretty much involved in that I just thought of it when Dr. Campbell mentioned subjective measures. I think there is one case where there probably isn't any subjectivity in the measurement procedure, but there may be questions about validity of the data that you're collecting. Jack, I don't know if there's anything you want to say about that, but . . .

Capt Thorpe: I kind of disagree. The reason for that is that we develop a lot of interaction between the computer and flying in the simulator and you can get a guy in there and the students can thrash around and you can collect 35 parameters 20 times a second and come out with reams of data. If you're skillful you can reduce that to even one number like 25% of the time he did well or something, but then when you go to the back room to find out how all these measures were devised, there's our pilot back there, with one of our skilled programmers, and he's figuring out what measures to measure. And I think it's pretty much the same judgment. It's his judgment of what he thinks are the things that should be measured, so maybe we measure them more reliably or more accurately or in a time domain that's much more specifiable. We measure perturbations that normally we might not be able to write down fast enough as they happen, but what it all boils down to is there's a great deal of subjectiveness in objectivity.

Dr. Guion: Last night in the bar we were having an intellectual discussion with Jack and he was describing what seemed to me to be the ideal measure for the evaluation of trainee pilots. This was brought about because of a test that I told him about that was used by the City of Honolulu to evaluate candidates for fire truck drivers, which involved a fellow named Martin Luke riding with the candidate up Tantalus Road. If any of you have ever been near Honolulu and know what Tantalus Road is, just visualize taking that road with a full length fire truck. The score, Martin would say, was the number of drinks he had to have before he could write up a report after he got back down to the valley. Jack's story was that the real score would be measured by the pressure of the check pilot's hands on the arms of his seat. Now all of that, of course, is barroom nonsense, and like a lot of other barroom nonsense there's a great deal of wisdom in it. Obviously, the bar-seeking check rider with the fire truck or the arm gripping check pilot is making a subjective judgment about the quality of the performance of the person being scored with either of these. The question now becomes one of how you record that subjective judgment and I submit that a 5-point rating scale is not going to be as reliable and valid a recording of the subjective judgment as some kind of a dynamometer on the arm of that chair. And of course what you'd have to do is develop some sort of a personal equation, a kind of a chicken factor, for different observers so that you could make a correction for the timid versus the foolhardy check pilots. But the point is that even though it is still a subjective evaluation you're getting here, your method of recording that evaluation does not always have to be a rating scale. And I think we ought to investigate some other approaches to recording subjectivity. I don't see any good reason why you can't do a little selling with these guys that do the check riding and get some GSR data if nothing else and use it as a criterion measure. And I am being only maybe 10% facetious.

Col Ratliff: I might point out that the Russians have a little test called the "Falling Down Test" that they use in their pilot selection program with which they instrument the individual's blood pressure and pulse and things like that, and all he has to do is stand up straight and fall forward on the floor. And the intensity of his physiological reaction during that period was recorded and held against him, I presume.

SUMMARY STATEMENTS

Dr. Guion: I'd like to say first of all that this has been a marvelous vacation. I've been here now nearly two days and the Federal Courts have not really intruded themselves into the discussion at any point. And I do not recall a 2-day period, other than when I was on vacation, when I have been free of concern for the effects of courts in the last year. I'm not entirely sure that you should have given me that vacation, because I think that some of the concerns for the courts may become your concerns. And even if they don't, one of the effects of the court involvements that we've had over the last few years has been a re-thinking, a very needed re-thinking, of the whole concept of employee selection, validation procedures, etc. And I think that you could do well to raise the same kind of question with all of the things you're doing, namely, how would I defend this if it were challenged in court.

I raise this question particularly with regard to this material that has been given the unfortunate name by you people of synthetic criteria. This is one step worse, I think, from a semantic point of view, than synthetic validity, of which I have been guilty. Obviously, we are not synthesizing either validity or criteria. You already spoke this morning of the semantic absurdity of the phrase, but it means something more than that. It means that when we are not thinking about the courts, but are thinking only of our professional colleagues, we try to put everything that we do into a framework of validity whether it fits there or not.

Now if you look at the APA AERA NCME standards, you'll find that validity comes under three guises: criterion-related validity, construct validity, and content validity, which in a paper a couple of months ago I said doesn't exist. I think it's down to two. Criterion related validity is a pretty straightforward kind of thing except that it's not really concerned with the validity of a test, it's concerned with the validity of a hypothesis, the hypothesis that some measure can be predicted by some other measure, either actually predicted over time or in a purely statistical sense in a concurrent kind of study. Construct validity is a very complex idea and very few of the things that have been thrown about in court discussions of late under the heading of construct validity have any resemblance to the kind of Cronbach and Meehl notion of construct validity that started the notion several years ago. The point of all this is that in a Supreme Court decision there was a term used called "job relatedness." I don't know who developed the term, but I like to think of it as a legal term rather than as a psychometric term. And when we're thinking in terms of court involvement we have to identify arguments for convincing a court that a method of selecting people, whether it's a test or anything else, is related to performance on the job.

I would like to urge that all the work now being done under the heading of synthetic criteria be done simply under the heading of a systematic method for gathering judgments (see, I'm on that same kick even though it started out like I was talking about something else) a systematic attempt for obtaining judgments about job relatedness.

Now if we're going to talk about predictive validity, I think I'd like to point out that the proper aim of personnel research, whether it's selection or training or whatever, is to predict and influence future behavior or the consequences of future behavior. The purpose of personnel research is not to evaluate instruments. We can have a lot of fun designing studies to evaluate tests or training methods or something of this sort--but personnel research, even if it's sponsored by OSR, is primarily concerned with making more proficient personnel. This is its fundamental purpose and we can't lose sight of it.

Now, in that Army paper (and I'm using this statement, incidentally, as an illustration--not as a criticism--because I have no intention to criticize the intent of the paper, only the language), there's a statement that grades are used as criteria for cognitive predictors. I think that statement illustrates the backward way that we often think about personnel research. We have in our hands now a cognitive predictor; therefore, let us look for grades as a good criterion that we might be able to predict with this cognitive predictor. That is not our business. Our business is to say (a) we want to predict performance in training, or (b) we want to be able to predict proficiency on a job, or (c) we want to predict how fast people will reach some stated level of proficiency, or whatever, and then figure out what the best way is to predict that particular criterion. Ooops, I slipped. That was the second stage. The first stage is how to measure it. See, I'm disagreeing, Dr. Muckler, on your sequence. I think the value judgments should precede the measurement, not follow it as a transform. I'm trying here to make a defensive comment about the quotation that was attributed to me yesterday about "a criterion is simply something that we predict" because I'm trying to give the indication that the something to predict must be identified before we develop a measure for it. I'm also being a little bit defensive about the comment made by Dr. Mullins that it somehow jars to talk about the validity of a criterion measure. In the first place, I don't think that's really consistent with the rest of the paper, because if a criterion really isn't different from the predictor except in the point-of-time scale, then all of the validation that you would do with the predictor applies to the criterion measure too. I am not the least bit jarred by concern for the validity of a criterion measure. I get jarred when there isn't any such concern.

I guess the only other thing I want to say in the summary here is to reinforce the Navy's views, or at least Dr. Muckler's views, whichever they are, on simple versus multiple criteria. I think, and I wish that Mr. Camm were still here, I think it was rather shocking to find that these skill qualifications tests come up with a single score. These are complex areas of performance. There is no good reason to suspect that any one test is going to be predictive of all areas of performance in a skill qualification criterion, or that a job must necessarily always be done in precisely the same sequence or same manner. And I think that we need to move away from World War I and the implications of the straightforward time-versus-level kind of table into a recognition that those interactions that scared Ray so much yesterday are quite possible, and even if they don't serve as interactions, they may very well serve as additional main effects. I think that we have to pay a great deal more attention to the complexities of performance than can be carried out with some kind of a single number that is supposed to somehow represent a compensatory summary of all of the components that go into that number.

Dr. R. Campbell: One has to be courageous to hold a 2-day meeting on job performance evaluation and criterion problems. Most of the papers stuck somehow to the rubric, although we did seem to cover an awful lot of ground that wasn't focused on those two subjects.

There was some confusion in the papers, I felt, over the term "job performance," and I think that needs clarification. For my money, job performance refers to on-line behavior and output--what the person is actually doing on the job and producing. And I'd like to keep that pretty clean; that's what I mean by job performance.

Criterion can mean all manner of things and is not so specifically defined in my taxonomy. There was much discussion of purpose, and how important it is that we keep purpose in mind in selecting a criterion. I want to echo that statement and emphasize that we ought to keep these multiple purposes in mind when we're selecting criteria and when we're evaluating what we're doing.

It is essential that we be explicit about the distinction between proficiency and performance. Most of the predictors that we've been talking about, most of the selection instruments, in the last 2 days deal with proficiency, which hopefully will be related to a person's output and behavior on the job. Of course it's not hard at all to conceive of highly proficient people being not very good job performers. I think we kind of buried that along the way. If you bring me into a special setting where I have to be able to perform some maintenance function, it may be that I can perform that maintenance function better than most everybody else you can bring in, but you put me out on a job and I don't do very well. Okay, it's the old saw, it's what a person can do versus what they actually do on the job. Now if there's a disparity between the two, what a person can do and what they actually do on the job, I don't tie all that to motivation. We ought to look at management practices, which someone around here did mention. There was some agonizing that you haven't increased your validity coefficients in the last 10 years. I don't consider that an indictment. Maybe it's a function of what you're trying to predict. Perhaps you're somewhere near the maximum level of prediction just looking at selection instruments. And perhaps the focus must be broadened beyond selection.

Another purpose for validation or for criterion selection running through the papers was the acceptability of the criterion, which translated to the ability to sell our work. Acceptability is important in selection of a criterion and we must consider the user; however, I just raise a caution that we don't let the selling determine the research and that we lose our way in the process. And there are other purposes. What I'm trying to say is that while we recognize the importance of purpose, I'm not sure we always explicitly deal with purpose and let it guide what we're doing.

There were, for me, some other fuzzy definitional issues, but they've already been discussed. I come down strongly on the side of "the criterion problem is not just a measurement problem." It certainly involves values and judgment, and I just want to second that. I also liked the comment we heard that we get overly upset with complicated problems, and we ought to recognize that we're dealing with a very complicated area.

There was a statement at the outset about whether there is a glorious solution, and I can confidently say "No," because I don't think anybody's going to get one very soon. There is no glorious solution. I think you work, work very hard, at devising the best criterion feasible in a given situation. An example of this would be Christal's switching to a time criterion. There is no one solution to criterion problems. You react to the realities and complexities of the situation. I want to cite several things I particularly liked in the papers. One was the discussion of levels of criteria--from individual levels up through system level. I liked the emphasis on measures of proficiency and, although I don't know these systems very well, I fully support the intent of them--the real training, the symbolic performance testing, and the skill qualification testing. These kinds of proficiency measures should be very useful.

The major omission I noticed in the program was the failure to deal with job performance--the outputs and behavior of the person on the job. That is, aside from ratings of performance. I know this is a very difficult problem, it's a very expensive problem sometimes to try to fix, but, again, if we're looking for criteria, particularly for research purposes, I was hopeful that I would have heard more in the way of conceptualizing how one might get performance measures and the methodology that might be used. And I'm not so pessimistic as to say it cannot be done. I wish I had heard more about that.

Dr. Helmick: Well, I certainly won't be sufficiently presumptive to give an indication I'm going to summarize everything that happened. I'll say that I think I've agreed with the summaries to date, and I'm sure I'll agree with those to come. I have down here a note from this morning's presentation by Col Ratliff that certainly ties in with what Bob Guion said, and I marked it important, underlined, "improved way of making judgments." I think that really gets to the heart of much of what we need to be dealing with.

I don't know that I am disagreeing really with Dr. Campbell on making the sponsor or the client happy. It's pretty clear that one can overstress that. I think on the other hand from my own experience in applied work and from earlier military experience, it can certainly amount to an awful lot of wheel spinning if you don't have some agreement or some understanding as to what it is that's going to be acceptable and usable. In an entirely different context I recently picked up a phrase, "Oh, yes, there's a need but there's not a want." And until that want is recognized, perhaps created, the need may be irrelevant.

I do want to congratulate the group for, first of all, recognizing the problem. I think nobody expected that we would have all the answers at the end of 2 days, but I think it has been, to me at least, a very useful discussion and it's very gratifying to me to see the approach and the attack that's being taken. There are two or three things that I think need to be given some attention. A number of you, I suspect, heard Harold Gulliksen's invited address at the last APA meeting. The approach he was describing at that time is really one of reversing the predictor-criterion priority and recognizing that sometimes when you get low relationships between predictors and criteria the answer may very well be to examine the criterion because you frequently know much more about the predictor; you have a much better understanding of what it really is than you do for the criterion. The classic example of course is the one that he used: the early Navy experience in which much to some of the psychologists' surprise when they began to validate the Navy classification test against performance in training, the discovery that for a Naval Machinist Mate, the highest validity was for a verbal test and one of the lowest was mechanical aptitude, and yet the task was very clearly primarily that of a mechanic. Well, the simple solution, of course, would have been to accept the criterion and to utilize the obtained validities as a method of selecting people who would do well in the course. But fortunately somebody said this really doesn't make too much sense, let's look at the criterion, which was course grades. And the grades were on a written examination based entirely on lectures and textbook material. And the individuals in the course said they never had their hands on anything that resembled a piece of armament. So this analysis led to developing what was called the Breech Block Assembly Test which involved actual disassembly and reassembly of a mock-up of a part of a naval gun. And lo and behold when that was used as a criterion for success in the course, the mechanical test had a fair amount of validity and the verbal aptitude test dropped considerably. I think the principle which it illustrates is that you can sometimes get a great deal of information about the criterion, or you can at least raise very meaningful questions about the criterion, by looking at the relationship that known predictors have with it.

A somewhat related topic is that of the effect of the criterion on the training. Again, this is not really aimed at solving the criterion problem, but in dealing with the criterion problem I think one has to recognize that the criterion may become a primary determiner of training, or at least of learning from the standpoint of the student. The naval example I just gave may be a case in point. In our own area the example we always come back to is that of essay testing versus objective testing for writing (composition). I think both the College Board and ETS, at least the vast majority of the staff, would take a very strong position that from a measurement standpoint with a given period of time and a given cost, there is no reason to use anything other than objective measures of writing for measurement of writing ability. On the other hand, it is pretty clear that as long as the objective test, the marking of blanks on the answer sheet, is the only thing that seems to be being evaluated, it's pretty hard for teachers to spend the time in grading and collecting essays, and it's pretty hard to convince students that they should do anything other than learn some of the techniques of taking objective tests. So from time to time the College Board has been convinced that, not for measurement reasons but for educational purposes, the criterion has to take the form of actually getting students to do some writing. In some of my overseas experiences in looking at the situations in other countries, it's very clear that the criterion, the final examination, so determines the curriculum that in many cases the purposes of education are rather completely subverted. So all I'm saying is that in dealing with the criterion, its effect on the whole learning and training process needs to be kept in mind.

One last point has probably been implied in all that has been said. I think there's some tendency to ignore basic considerations of reliability of the criterion if it seems to be objective, if it seems to be quantified, if it seems to be specific, and this is not necessarily enough. And here I go back to my World War II bombardier research experience where in all of the three flying training categories, bombardier, pilot, and navigator, the criterion that seemed to be the best and the most objective one, where you can really get numbers that almost popped out at you, was the average error of the students on bomb drops in training missions. And it was really a rather horrifying discovery when people came up with the fact that you couldn't predict the average error on the odd missions from the average error on the even missions. The reliability was essentially zero for as objective, as quantified a criterion as one could find. We managed to do a slight follow-up on this. What started out to be a very closely controlled experimental class was, fortunately, disrupted by the Japanese surrender. We got enough evidence, however, to indicate that, in terms of very carefully measured bomb dropping performance, the least important link in the whole chain was the bombardier. The airplane, the auto pilot, the degree of turbulence that day, the actual pilot flying the plane, the bombsight, all of these things, according to an analysis of variance, contributed more to the average error than did the bombardier. So we need to take a hard look at criteria even though they seem to have the highest possible face validity and not be lulled into any sense of false confidence.

Dr. McCormick: I don't think anything in the papers we have heard here in these past couple of days could be viewed as a quantum step in the area of criterion development, but I think there are some overtones that do warrant some recognition and that offer at least modest encouragement for the future. In the first place, I believe I sense a seriousness of concern about this problem in the military services that hopefully will provide the momentum for concentrated attention on this problem which is clearly a critical one in connection with personnel research. In the second place, I believe there are a few bits of wheat mixed in with the chaff that might take roots and develop into some new strain of criteria or approaches to criterion development.

Although we intend to seek the Holy Grail of the ultimate criterion of job performance, we certainly should not bypass the operational need for criteria of achievement in training as referred to by Meyer in his discussion of instructional development systems and as discussed by DeLeo in the paper by himself and Waters. I was quite impressed by Fred Muckler's paper in which he referred to the many facets of criteria from A to Z, or maybe from AAA to ZZZ. I suspect that he must have lain awake many nights to organize what I believe to be a very significant discussion of this problem, in particular in crystallizing a number of points and issues which have otherwise been lurking furtively in the background. I think especially his listing of the criteria of criteria is one that might well be posted on the walls of research offices in much the same manner that many homes used to have framed mottos on walls such as "God Bless Our Home."

In winding up there are just a couple of points I might add. In the first place, I would like to suggest some attention to the notion of quality control. This is not a new notion, although it has not been mentioned in the confab here. I think that quality control as applied to human performance evaluation is something that has some sort of relevance to the problems with which we deal. And the techniques and approaches of the industrial engineers in connection with quality control of physical products and processes is one that I think can well be applied to the performance of people on their jobs.

And next I will reflect an admitted bias in suggesting that I believe the military services should pursue the notion of what I prefer to call job component validity, previously called synthetic validity or generalized validity. This would require the development, for a good sized sample of jobs, of information about the relationship between job components on the one hand and the human characteristics of those performing the jobs on the other hand. Such an analysis might offer the possibility of applying the relationships so teased out to other jobs, thereby avoiding the necessity of developing criteria for each and every job classification.

In closing, I would like to say that I am really impressed by the sense of commitment of the individuals who have presented papers at this seminar in terms of their interest in the criterion problem and also some of the notions that have been bandied about. I would be surprised if, as a result of this seminar, there would be any really earth shaking results that would solve the criterion problem for all time. At the same time, one would hope that this seminar would at least result in the exchange of ideas regarding this important problem to the extent that some 3, or 5, or 10 years hence, one would be able to look back and say that the development in this area has been moved forward at least by a few steps because of the organization of this particular symposium.

Dr. J. Campbell: A lot of excellent material has been presented in the last 2 days and it's not easy to digest it all so soon. Also, members have stolen most of the thunder. However, let me begin by describing several general impressions stimulated by the discussion during the last 2 days.

One dominant impression that does strike me is one that I have always related to classes that I teach in industrial/organizational psychology. That is, if one considers the major groups of applied psychologists in the United States who deal with problems like this, the researchers who are the most sensitive to such problems and who seem to have the best grasp on their subtleties are the military psychologists. On the basis of what's happened at this conference, I don't see any reason to change that opinion.

A long time ago, in 1967, I went to a similar conference sponsored by the Richardson Foundation. It took place in North Carolina, it was attended by a number of industrial/organizational psychologists, and it was on the criterion problem. In comparing the discussions there with the discussions here, I must make the judgment that the field has come a long way, at least the level of conceptual understanding is much higher now than it was then.

In the same breath, I would like to say that the criterion problem, as we have historically talked about it in this field, is intractable. There is no solution to the "problem" and we should all get away from the notion that a final answer will someday present itself. However, one major reason the criterion problem is insolvable is because of the way we traditionally have defined it. For example, I would like to sentence Robert Thorndike to 40 years of computing factor matrices by hand for making the distinctions between immediate, intermediate, and ultimate criteria. The concept of the ultimate criterion has been the bane of our existence and it should be stricken from the language. There is no such thing. However, regardless of its label, we seem to have striven in past years for something. What is that something? My own guess of what's in everybody's mind is that it's one kind of rating or one kind of measurement that will be generalizable across all situations, at least in form if not in content, and which will almost always yield high reliability, relevance, and predictability. All this is in spite of the fact that it is very reasonable to conclude that performance in certain situations is at best not very predictable and at worst probably random, and that no one is ever going to find a predictable or even reliable criterion in such contexts. This is not the fault of psychology and it is not the fault of applied psychologists. It is simply the way certain jobs happen to evolve in certain kinds of organizations. We may want to think about changing the organization itself, so as to make performance more predictable; but adopting the goal of finding predictable measures, when no predictability or even reliability exists, has given us terrible guilt feelings, and we make almost pathological responses to the "problem" as a result. Therefore, one general conclusion I would like to make is that we really should redefine the criterion problem drastically and adopt a different way of thinking about it that does not include things such as I just mentioned.

Dr. Brokaw made a statement early on which I would like to reemphasize. He said that perhaps what we should be aiming for, if there's any one thing, is a useful strategy for arriving at criterion measures. That is, what we need is an overall plan for how to approach the development of a criterion, not a set of specifications for the criterion.

A number of the following points have been mentioned already, but I would like to consider them again briefly. First with regard to the problems of ratings, notice how easy it is to forget the parameters one should not forget. Guion reminded me pointedly that "Well, you must distinguish research versus operational kinds of measures." I did forget to do so, and I am sorry. Besides distinguishing between research criteria and operational criteria that are actually for the purpose of appraising people, there are also criteria that have as their main purpose maximizing the usefulness of performance feedback. That is, these are criteria that are appropriate for training and development purposes, but which are probably not very useful for research or appraisal. Just to state the obvious moral, it's very easy to forget the purposes for which we are going to use a specific measure. Such forgetfulness is an insidious disease. What is the best way to innoculate ourselves against it? I don't know, but we should keep trying.

Second, I would like to echo what Dick Campbell said about the military's work on performance simulations. It is pretty exciting stuff. Obviously when using simulations there are pitfalls that must be faced at some point. For example, if there are truly important decisions to be made about people on the basis of such a measure, then you might find the same phenomenon that Dr. Helmick just mentioned with regard to the educational setting. People will start emphasizing the behavior measured by simulation and not the job activities they had been concentraing on previously. That is, if people are to be rewarded for high scores on a particular kind of measure, that's what they will try to maximize. We simply can't get away from B.F. Skinner. If simulation continues to become a more widely used method for assessing performance, then we really must worry about whether it is the specific behaviors on the test that we want to emphasize.

Also, as I mentioned before, it is easy to slip into thinking that performance assessment, whether it be for research criteria or for any of the other purposes, takes place in a vacuum. Even if we make the argument that a particular study is for research purposes only, we still must worry about how people are cooperating and how they actually respond to what we ask of them. We seldom go back to people and say, "We asked you to participate in a research project, is that what you really thought was going on? How did you respond to the briefing? etc.?" Such an examination of the research process could be very informative, if pursued.

Regarding Guion's comments about the economics of criterion research, I would like to speak as one citizen (i.e., taxpayer). It really is disconcerting that people back "there" will waste so much money on so much else and then starve to death one of our most important military manpower problems.

One curious thing I noticed about the last day and a half is that the literature being cited wasn't very recent. I seldom heard a date beyond the late 60's. The 1970's were mentioned very infrequently. I'm wondering if that's because nothing's happened during that period, or it's not worth much, or what. It's just an impression I have, and perhaps it is inaccurate.

Finally, switching from describing impressions to giving advice, let me make two or three suggestions. One is that I agree with Dr. Muckler that we really have to stop sounding so pessimistic. We really know a lot more about criteria and the criterion problem than we give ourselves credit for and we ought to tell people that. We shouldn't keep making ourselves look so bad. For example, long before the discrimination question reared its head in the selection domain, we talked ourselves into the notion that we had to have perfect predictions in order to do our job right, or at least correlation coefficients of .75 or better. It is not surprising that when lawyers and the courts came along, they looked in our textbooks and assumed that near perfect prediction was possible, if only the psychologists would get their heads together. As a result, it is now our fault that prediction isn't perfect.

Something about which we didn't hear enough is that a criterion measure directly reflects the values of the organization concerning what individuals should be doing. By implication, those things which are selected as criteria are those things which the organization says are important for people to do. The criterion is the variable of real interest. Now, what is the value system of the organization? The obvious answer is that it is many things, and there are those within the organizations who disagree. For example, in your situation you might be putting together a measure of pilot proficiency and there could be wide disagreement within "management" as to whether a high score or a specific component of proficiency is good or bad. In any organization, if you carry out the criterion development process correctly, you will involve the users, management, and the rank and file, and concern yourself with trying to find out their values and preferences for how high and low performance should be defined. Such a process will most likely uncover serious conflict. Certainly that is the case in educational institutions like large universities. It is legitimate conflict about what behaviors and accomplishments are valuable to the organization and which are not. I think we have to program into our criterion development activities some more systematic procedure for confronting such value conflict and dealing with it.

One kind of research that I personally would like to see conducted more frequently by people in the military and elsewhere, has to do with more applied investigation of the judgment process itself. Dr. Guion mentioned a little while ago the notion of a "true" score on a performance dimension to be rated. Some associates of mine in Minneapolis (Borman, 1978), under the sponsorship of ARI, conducted a study in which they tried to program the performance of the people to be rated as precisely as they could. That is, people were given scenarios of behavior episodes that illustrated examples of high, medium, and low performances on various dimensions. By careful rehearsal of the "actors," the experimenters tried to establish a true score for performance at various levels. That is, they were trying to set up a situation where the performance of an individual on each of the factors was known. The questions to be investigated concern what the observers do with the performance information. Do they make large errors? Are errors systematic or random? What kind of sytematic errors are present? What method of assessment yields the smallest error? The behavior episodes used by Borman were rather brief, which perhaps was the study's main drawback; nevertheless, I think the paradigm could be applied in many different contexts. However, it should not be translated into broad survey format. We have had enough of that. The method would quickly lose its fidelity there. In sum, I think we could learn a lot about what's going on in the performance judgment situation by a more intensive look at the actual processes that take place.

Also, I don't think we've done enough with trying to develop better methods for sampling task behavior. We've too quickly jumped to a consideration of how well people can rate performance dimensions. I'd like to go back to more research on how best to do the actual sampling and describing.

Let me leave you with a very non-traditional question for our kind of applied psychologist. What would happen if we put on a B.F. Skinner, Inc. hard hat and routinely did an operant type functional analysis of every performance assessment situation that we encountered? What rewards and punishments control the behaviors of the subjects, and ultimately, the ratees? What rewards and punishments control the behavior of the sponsors of the research? In general, what are the reinforcers and the reinforcement contingencies that control the entire criterion development and performance measurement system? Without a clear understanding of such relationships, criterion research often must swim upstream. It would be better to go with the flow, so to speak.

IMPRESSIONS

COMMENTS ON SYMPOSIUM ON CRITERION DEVELOPMENT FOR JOB PERFORMANCE EVALUATION

John S. Helmick

I appreciated the opportunity to participate in the criterion symposium and found it a stimulating experience. I congratulate the Air Force for its recognition of the importance of this problem and its straightforward approach to trying to deal with it.

I also want to commend the sensitivity expressed by several individuals to the importance of sponsor acceptability and user requirements in making applied research effective. The communication between the researcher and the user seems to me to be one of the most critical features in being sure that applied research is actually applied. This should be a major concern from the initial definition of the problem to the preparation of the final report.

When to Measure the Criterion

Perhaps the major problem in the whole area is that of determining where in the time frame to attempt to define and measure the criterion. Should the criterion be a measure at some point in training or on the job? If the latter, should it be initial performance or later performance? Almost inevitably the accuracy of prediction decreases as the time span between its measurement and the criterion measurement increases, yet the importance of the criterion increases. While some of the intervening variables between prediction and later performance can be anticipated and accounted for in prediction, in general they cannot be. To deal with this we need a better understanding of the chain of events between initial and final measurement and as much knowledge as possible of their interrelationships. One of the papers suggested that the differentiation between predictor and criterion was essentially in the time at which each was measured and supported the procedure of successive measurement. While I am not willing to accept the principle that the criterion does not have a kind of meaning different from that of the predictor, I agree that attempts to differentiate them on some simple all encompassing basis lead to difficulties. The answer, if any, to the problem lies in understanding a network of successive measures taken throughout the time span and this really implies that one must determine and understand the underlying psychological principles that relate antecedent to consequent, if not actually cause to effect. While this ideally calls for longitudinal study, a series of short-term, almost cross-sectional, studies may yield satisfactory approximations. All this suggests that satisfactory criterion research may very well be basic psychological research and that work that does not recognize this may have significantly less general payoff in the long run. In this connection the distinction made between evaluation and measurement should be kept in mind. A satisfactory definition of criterion performance does require value judgment.

Restriction of Range

In addition to this somewhat philosophical general statement there are a number of brief comments that may be worth noting. A number of participants pointed out that the restriction of range after training does present a problem. This can become an even greater problem if the predictor measures are used as a basis for compensatory training. If effective this will produce a self-defeating prophecy making the predictor seem less useful than it really is. The suggestion should be pursued that one look for differences among individuals considered successful by the traditional criteria already in use. Some of these differences may provide a basis for new criterion development.

Reverse Validity

The concept of "reverse validity" is worth pursuing. This is simply the recognition that predictor measures are often better understood than the criteria, and hence high or low relationships may provide insight into the nature of the criteria. This contrasts with the usual procedure of accepting the criterion as the given and judging the predictors on that basis.

Unreliability

Considerable attention was given to the unreliability of criteria, particularly when they take the form of ratings, essays, oral examinations, and other admittedly subjective judgments. This is all to the good, but it should not allow one to assume that because quantitative, apparently objective, accurately measured performances are used that reliability will be automatic. The unreliability of error measures in practice bomb dropping during World War II cadet training is a case in point.

Effect on Training

In searching for and introducing criterion measures, it is important to be alert to their effect on the training process. This is especially true if the criterion involves sampling a relatively small number of all those items which should be included in training. In such a case the inevitable tendency is to gradually concentrate the training on only those items that are to be evaluated.

Speed

It seems worthwhile to pursue the measurement of speed as a way of dealing with the criterion. I see this as largely a matter of a different way of measuring the criterion rather than producing a truly different criterion. One still has to make the judgment about and decision on what behavior will determine the end point for measurement of time consumed.

Group Performance

It was recognized that some of the examples described really involved performance of a group or system rather than that of a single individual. It seems desirable to keep these two types of performance separate. While individual performance can frequently be aggregated to provide a group measure, it is likely that in many cases the group performance will require some separate measurement of group outcomes.

Cognitive Emphasis

One final note. I was struck with the continuing emphasis in many of the presentations on cognitive and intellectual variables. I would not want to underestimate their importance and, as one who's been involved in work almost entirely concerned with such variables for many years, I recognize the much greater ease of measuring them. Nevertheless I think we can continue to be lulled into false feelings of success by putting too much weight on such measures.

COMMENTS FROM THE SIDELINES

Ernest J. McCormick
Professor Emeritus
Purdue University

A pervading theme of this symposium has been the need for concentrated effort directed toward "doing something" about the criterion problem in personnel research in the military services, with particular focus on the measurement of on-the-job performance.

Although I think nothing in the papers we have heard could be viewed as a "quantum step" in the "solution" to this problem, I think there are overtones that I believe do warrant some recognition, and that offer at least modest encouragement for the future. In the first place, although I may be a bit Pollyannish, I believe I sense a seriousness of concern about this problem in the military services that hopefully will provide the momentum for concentrated attention on this problem, which is clearly a critical one in connection with personnel research. And in the second place, I believe there are a few bits of wheat mixed in with the chaff that might take root and develop into some new "strain" of criteria or approaches to criterion development. Let me now touch perhaps a bit randomly on a few of the points that were made in the papers that seem to me to be of particular interest.

To begin with, Mullins and Ratliff in their discussion of the "Criterion Problems" emphasize the point that the best predictor of future achievement is some indication of past achievement. (This theme, of course, has been expressed by various people, including Wernimont and Campbell in their paper "Signs, Samples, and Criteria.") Following along this line, they raise the question as to whether there is really any difference between predictors and criteria, since both are measures of achievement of some type. I have a great deal of sympathy with this point of view, since predictors are measures of some type of achievement. But granting the basic thesis that predictors and criteria both are measures of achievement--that is, that they do not differ in their natures--I believe that at least in many circumstances they do differ substantially in "degree," particularly the degree of complexity. In other words, I believe that criteria generally are harder nuts to crack than predictors.

I was interested in the listing by Weeks and Mullins (in their paper on "Rater Accuracy") of the basic dimensions of the rating paradigm, these being: (1) the rater, (2) the person rated, (3) the traits or tasks to be rated, (4) the social environment, and (5) the physical environment. I believe all of these warrant systematic investigation as sources of possible variance (error and otherwise) in criterion development, and I certainly support their proposal to explore some of the problems associated with raters. It would be particularly useful to be able to identify those individuals who can serve as good raters and also to explore the extent to which training of raters can improve their performances. (In a study we have just completed it was found that even moderate training of raters had some beneficial effect upon the ratings made by them.)

Aside from the factors which they mentioned, however, I believe there is another area that warrants substantial attention, and that relates to the type of "rating" procedure that is used. Curton, Ratliff, and Mullins in their paper "Content Analysis of Rating Criteria" do in fact refer to this matter, in particular by referring to the use of behaviorally anchored scales as contrasted with conventional rating scales. However, there may be other approaches to the development of criteria that might also be subject to some comparative analysis. Other types of rating procedures of course include the forced choice method, the weighted checklist, the various personnel comparison systems (such as rank order, paired comparison, and forced distribution), and the critical incident technique. Actually most of these methods of obtaining personnel appraisals differ in the nature of the human responses that are required. For example, the forced choice checklist and the weighted checklist depend pretty much on the "description" of behaviors rather than making evaluative judgments about behavior. In turn, the conventional rating scale requires the making of absolute judgments as contrasted with the relative judgments that are required by the personnel comparison systems. Various experimental studies dealing with judgments people can make about physical stimuli indicate that people are much better in making relative judgments than in making absolute judgments. I thoroughly support Christal's argument for the use of relative judgments in at least many circumstances where human judgments must be used.

Along this line, the suggestion made by Mullins and Weeks in their paper "The Normative Use of Ipsative Ratings" is a rather intriguing one. In addition, I might refer to some of the notions that were suggested back a few years ago in this same hotel relating to performance evaluation of Air Force officers. A number of rather ingenious suggestions were made at that time that might be further explored in connection with their relevance in developing criteria.

Thus, I would urge further comparative studies of different methods of evaluating the performance of individuals, including comparisons of different "types" of human judgments, both as related to their psychometric properties and to their practical differences, as referred to by Curton, Ratliff, and Mullins.

However, we must recognize that no rating procedure is going to compensate for poor human judgments. Granting this ubiquitous fact, however, we should try to find out as much as we can about the processes of making human judgments, toward the end of the development of the "rating procedures" that provide the best opportunity for eliciting the best judgments people can make. Bob Guion's public pronouncement of a mid-career shift to investigate the processes of making human judgments is indeed an encouraging sign.

The development and use of on the job sample tests or what Foley refers to as performance measurement (PM) certainly deserves some place in the military system. There is probably no question but that the use of such tests can provide a reasonably adequate basis for the derivation of criterion values and performance measures of individuals. The problem with respect to such measures, as we all know, is that of time and cost. I presume the basic problem here is one of somehow determining those areas and types of jobs for which this time and cost would be cost effective if such tests were used, as contrasted to those areas where it would not be cost effective. Because of the practical problems of cost and time involved in development and use of job sample tests, however, I would urge further exploration of the "simulation" of such tests as Foley suggested, and of the extent to which "sampling" the performance of various aspects of the job can produce criterion values that may approximate measures of performance on the total job. The theme of simulation and sampling was also referred to by Mullins, Ratliff, and Earles in their paper on "Synthetic Criteria." If they confirm the finding that their "R-technique" and "M-technique" provide the basis for deriving estimates of performance that are strongly correlated with actual performance, such a route is one that should be pursued.

Although we tend to seek the "Holy Grail" of the ultimate criteria of performance, we certainly should not bypass the operational need for criteria of achievement in training, as referred to by Meyer in his discussion of Instructional Development Systems (IDS), and as discussed by DeLeo in the paper by him and Waters.

I was very much interested in Dr. Christal's suggestion regarding the use of "time" as a criterion, with variations in terms of the speed of acquisition, decay, and reacquisition. I personally feel that the time taken to learn something is, at least on rational grounds, an indication of learning ability, and feel that efforts to use time as the basis for establishment of criteria might well warrant considerable attention. In this regard, however, I might comment on one possible problem, and that is the problem of determining the point in time at which a person's performance or acquisition of skill has achieved a "satisfactory" level. (Although time might then be a relevant measure, the use of this does not completely avoid the need to make some determination as to the "level" of performance of individuals.)

As a sideline comment about time, I might add another point, that the stage at which a person initiates his learning presumably is an important factor in the time taken to achieve some previously determined level of proficiency. This matter has been rather thoroughly explored by Stanley Lippert, to the point that he has derived an "equation" for taking into account the stage of skill at which the person starts training, and has found that this improves very significantly the prediction of the future learning of the individual.

I was quite entranced by Fred Muckler's paper in which he covered the many facets of criteria from A to Z, or perhaps from AAA to ZZZ. I suspect he must have lain awake many nights pulling together and organizing what I believe to be a very significant discussion of this problem, in particular in crystallizing a number of points and issues which have otherwise been lurking furtively in the background. I think especially his listing of criteria for criteria is one that might well be posted on the walls of research offices in much the same manner that many homes used to have framed mottos on the wall such as "God Bless Our Home."

Before I close I would like to add three additional reflections. In the first place, although ratings have been thoroughly maligned many times over (and certainly with some justification), there are at least a couple of factors that will probably cause them to be with us for a long time. There are some aspects of human behavior for which human judgments probably are the most appropriate basis for evaluating performance. Furthermore, there are some aspects of human performance that conceivably should be evaluated on the basis of some "objective" measures--but for which we have not been bright enough to figure out adequate methods of measurement. In such instances the basic problem may be one of figuring out the best way of obtaining reliable and valid judgments, rather than being overly obsessed with the notion of obtaining "objective" measures of performance.

In the second place, I would like to suggest further attention to the notion of "quality control" as applied to human performance evaluation. This is not a new idea, of course, but I believe it has some further relevance to the criterion problem.

And in the third place (in which I will reflect an admitted bias), I believe the military service should pursue the notion of what I prefer to call job component validity (previously called synthetic validity). The development, for a good sized sample of jobs, of a solid data base characterizing the relationship between job components on the one hand, and human requirements for performing the activities involved in them on the other hand, might offer the possibility of applying the relationships so teased out to other jobs, thereby avoiding the necessity of developing criteria for each and every job classification.

In closing I will say that I am really impressed by the sense of commitment of the individuals who have presented papers at this seminar, in terms of their interest in the criterion problem and also at some of the notions that have been bandied about. I would be surprised if, as a result of this seminar, there would be any real earth-shaking results that would "solve" the criterion problem for all time. At the same time, one would hope that this seminar would at least result in the exchange of ideas regarding this important problem to the extent that some 3, or 5, or 10 years hence one would be able to look back and say that development in this area has moved forward by at least a few steps since this time.

☆U.S. GOVERNMENT PRINTING OFFICE: 1979-671-056/24